UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Caracterización de la divulgación de ciencia ciudadana en Twitter mediante análisis de redes sociales y minería de textos

Máster Universitario en Ingeniería Informática

Autor: RIBAS GÓMEZ, Diego

Tutor: HAYA COLL, Pablo Alfonso

Departamento de Ingeniería Informática

Agosto, 2020

Caracterización de la divulgación de ciencia ciudadana en Twitter mediante análisis de redes sociales y minería de textos

Máster Universitario en Ingeniería Informática

Autor: RIBAS GÓMEZ, Diego

Tutor: HAYA COLL, Pablo Alfonso

Departamento de Ingeniería Informática

Agosto, 2020

Agradecimientos

La vida no es, sin duda, un camino sencillo, y en muchas ocasiones nos sentimos perdidos y desorientados. A veces la vida nos golpea y nos desviamos, pero por suerte, con tiempo y con la ayuda de las personas que nos rodean, conseguimos seguir caminando y retomar el sendero. Este agradecimiento va dedicado a todos aquellos que me han acompañado y guiado para encontrar de nuevo el camino, haciendo posible este trabajo.

En primer lugar, agradecerle a mi familia, que son quienes me acompañan desde el inicio y quienes lo harán hasta el final. A mis padres por su apoyo, su preocupación, y su empeño en que no desistiera y finalizase el máster. Espero que se sientan tan orgullosos de que lo consiga como yo lo estoy de ellos. A mis hermanos, por haber hecho tan fácil crecer y aprender juntos, pero también los unos de los otros. Por su ayuda y por demostrarme que en todo momento he podido contar con ellos y que siempre lo podré hacer.

A mis amigos, compañeros de aventuras en lo bueno y en lo malo, por explorar conmigo y amenizar la ruta.

A Mercedes, por ser mi oasis en el desierto y darle un nuevo sentido al camino.

Y para finalizar, me gustaría destacar y agradecer especialmente a mi guía durante este tramo, mi director Pablo Alfonso Haya Coll, por haber confiado en mí y por su infinito compromiso y dedicación, habiéndose mostrado siempre disponible para orientarme y ayudarme con este proyecto.

Resumen

El objetivo del Trabajo Fin de Máster (TFM) es caracterizar los proyectos de ciencia ciudadana que se están realizando en la actualidad a partir de los datos que estos publican en la red social Twitter. A partir de esta caracterización se espera que emerjan los proyectos de ciencia ciudadana más importantes, las relaciones que puedan existir entre ellos, y las áreas de conocimiento donde más están impactando estos proyectos. El resultado permitirá determinar, además, qué organizaciones, e investigadores están impulsando estas iniciativas.

En esta línea, se ha implementado un sistema informático que realiza el análisis de las conversaciones que se producen en un Twitter. El sistema extrae grupos de conversación mediante algoritmos de detección de comunidades, analiza la actividad de publicación y los temas para cada comunidad, y realiza una adecuada visualización de los resultados accesible en línea. Se ha partido de un conjunto de datos en formato en CSV que recoge todos los tuits publicados desde principio del año 2020 en temas relativos a ciencia ciudadana.

Palabras clave: ciencia ciudadana, redes sociales, minería de textos, minería de grafos, Twitter, Python

Abstract

This master's thesis aims to characterize current citizen science projects from the data which they publish on Twitter. This characterization can be used to identify the most relevant projects, possible connections among them, and the areas of knowledge in which they have a bigger impact. The outcome of this analysis will also determine which institutions and researchers are promoting these initiatives.

For this purpose, a software to analyze conversations on Twitter was developed. The system extracts conversation clusters using community detection algorithms, analyzes the activity and topics of each community, and produces graphic representations that are available online. The starting dataset is a CSV file containing all tweets regarding citizen science published since the beginning of 2020.

Keywords: citizen science, social networks, text mining, graph mining, Twitter, Python

Índice de contenido

Agradecimientos III
Resumen V
Abstract VI
1. Introducción 1
1.1. Motivación 1
1.2. Objetivos 2
1.3. Alcance 2
1.4. Estructura del documento 3
2. Metodología 4
2.1. Recopilación de datos 5
2.2. Ingesta de datos 5
2.3. Preprocesamiento de datos 6
2.4. Carga y procesamiento de los datos 7
2.4.1. Carga de datos 7
2.4.2. Creación de grafos 8
2.4.3. Minería de textos 11
2.4.4. Detección de comunidades 14
2.4.5. Información sobre las comunidades 18
2.5. Presentación de los datos 19
3. Resultados 20
3.1. Resumen de los datos procesados 20
3.2. Creación de grafos 20
3.3. Minería de textos 23
3.4. Detección de comunidades 27
3.5. Información sobre las comunidades 30
3.6. Presentación de los datos 31
4. Discusión 35
5. Conclusiones y trabajo futuro 38
5.1. Conclusiones 38
5.2. Trabajo futuro 38
Referencias 40

Bibliografía 41
Anexo I: Código fuente y ficheros 42
Anexo II: Grafos completos 43
Anexo III: Distribuciones de grado 44

Índice de figuras

Figura 2-1 Esquema de la metodología 4
Figura 2-2 Ejemplo cabecera y filas del fichero 6
Figura 2-3 Ejemplo de grafo de retuits 10
Figura 2-4 Ejemplo de grafo de citas 10
Figura 2-5 Ejemplo de grafo de hashtags relacionados 11
Figura 2-6 Ejemplo de palabras con más ocurrencias 13
Figura 2-7 Ejemplo de evolución temporal hashtags retuiteados 14
Figura 2-8 Ejemplo de comunidades detectadas mediante Edge betweenness 15
Figura 2-9 Ejemplo de comunidades detectadas mediante Walktrap con pesos 17
Figura 3-1 Componente gigante del grafo de retuits 21
Figura 3-2 Componente gigante del grafo de citas 21
Figura 3-3 Componente gigante grafo combinado 21
Figura 3-4 Grafo hashtags 22
Figura 3-5 Hashtags con más retuits 23
Figura 3-6 Hashtags principales más utilizados 23
Figura 3-7 Top hashtags excluyendo aquellos utilizados por el bot 24
Figura 3-8 Hashtags totales con más ocurrencias 24
Figura 3-9 Términos con más ocurrencias 25
Figura 3-10 Evolución temporal de los hashtags más retuiteados 25
Figura 3-11 Evolución temporal de los hashtags más utilizados 26
Figura 3-12 Evolución temporal de los siguientes hashtags más retuiteados 26
Figura 3-13 Comunidades detectadas mediante Edge betweenness 27
Figura 3-14 Comunidades detectadas mediante Walktrap con pesos 28
Figura 3-15 Comunidades detectadas mediante Label Propagation con pesos 28
Figura 3-16 Comunidades detectadas mediante Walktrap sin pesos 29
Figura 3-17 Comunidades detectadas mediante Label Propagation sin pesos 29
Figura 3-18 Cabecera del sitio web 32
Figura 3-19 Boceto de la sección de contadores 32
Figura 3-20 Menú del panel de control 33
Figura 3-21 Ejemplo de sección con imágenes 33
Figura 3-22 Boceto de la sección de comunidades 34
Figura I-1 Estructura del repositorio en GitHub 42
Figura II-1 Grafo completo de tuits 43
Figura II-2 Gráfica completo de citas 43
Figura II-3 Gráfica distribución de grado en retuits 44
Figura II-4 Gráfica distribución de grado en citas 45
Figura II-5 Distribución de grado del grafo combinado 47

Índice de tablas

Tabla 3-1 Resumen de los datos procesados 20
Tabla 3-2 Información sobre grafos 22
Tabla 3-3 Comunidades en función del algoritmo 30
Tabla 3-4 Número total de entidades y proyectos en Walktrap 30
Tabla 3-5 Organizaciones, personas y proyectos con más ocurrencias 31
Tabla III-1 Distribución de grado del grafo de retuits 44
Tabla III-2 Distribución de grado del grafo de citas 45
Tabla III-3 Distribución de grado en grafo combinado 46

1. Introducción

1.1. Motivación

Se entiende por **ciencia ciudadana** a la investigación científica que cuenta con la implicación activa del público no especializado junto con científicos y profesionales. Formalmente, la ciencia ciudadana ha sido definida como "la recopilación y análisis sistemático de datos, el desarrollo de la tecnología, las pruebas de los fenómenos naturales, y la difusión de estas actividades por los investigadores sobre una base principalmente vocacional".

Una prioridad clave y un prerrequisito para asegurar el futuro de la economía y la sociedad europeas es aumentar el interés en la ciencia y la tecnología, así como mejorar el nivel de conocimientos científicos entre los ciudadanos, especialmente la generación más joven. Esto pone el foco en las iniciativas de ciencia ciudadana, que incluyen múltiples disciplinas científicas (ecología, climatología, biología, física…) y tecnológicas (fablabs y Maker Spaces), y que constituyen espacios abiertos para la construcción de conocimiento científico. Estas actividades requieren y generan conocimientos de la ciencia y la tecnología de manera que surja de la participación ciudadana y la autoorganización. Estas iniciativas son fuerzas motrices para aumentar la conciencia de la ciencia y fomentar la alfabetización científica en nuestras sociedades.

Sin embargo, aunque la base de nuestra economía y nuestra forma de vida se fundamenta en las aplicaciones de la ciencia y las nuevas tecnologías, hay una tendencia a que la confianza en la ciencia por parte de los ciudadanos está disminuyendo. En su lugar, las "verdades alternativas" como la negación del clima o los mitos sobre los riesgos de la vacunación ganan terreno incluso entre los ciudadanos bien educados. En este sentido, es preciso atraer a los ciudadanos a actividades científicas genuinas que fomenten su apreciación de la ciencia en la sociedad mediante la participación activa. La participación activa significa aprender haciendo a través de construyendo resultados tangibles (artefactos) y recolectando datos científicos relevantes para sus vidas. Estas dos dimensiones apoyan la adquisición de los conocimientos científicos pertinentes necesarios para la vida de los ciudadanos hasta la fecha.

Así, los proyectos de ciencia ciudadana permiten que el público, mediante la experiencia propia, comprenda la forma en que se conducen las investigaciones científicas. Muchos participantes descubren que el proceso de *hacer la ciencia* surge de la observación, de los métodos para la toma de datos y de las reflexiones o conclusiones a que estos conducen. Los voluntarios contribuyen fundamentalmente en la toma de datos, aunque también pueden completar investigaciones guiadas. Estos proyectos generan sinergias entre científicos y el público general formando equipos de trabajos mixtos.

Por dar algunos ejemplos de proyectos de ciencia ciudadana puede referirse a actividades como la recogida de datos meteorológicos, la localización de especies de

animales, el descifrado de documentos históricos escritos a mano, la resolución de puzzles científicos o la realización de experimentos en el jardín, pero también la formulación de preguntas de investigación e incluso al establecimiento de agendas de investigación, o la realización de proyectos científicos en colegios.

Actualmente existen multitud de proyectos que entran dentro de esta categoría distribuidos en todo el mundo, en múltiples áreas de conocimiento, e impulsados por diversas instituciones y organizaciones.

1.2. Objetivos

Este TFM está relacionado con los objetivos del proyecto europeo CSTrack (https://cstrack.eu/) que persigue desarrollar y aplicar diversos instrumentos innovadores en el estudio, la vigilancia y el apoyo a las actividades de la ciencia ciudadana. En particular, compartimos el mismo enfoque "observacional" para estudiar y caracterizar las actividades de la ciencia ciudadana basado en sus manifestaciones y rastros en la web y en los medios de comunicación social.

El objetivo del Trabajo Fin de Máster (TFM) es caracterizar los proyectos de ciencia ciudadana que se están realizando en la actualidad a partir de los datos que estos publican en la red social Twitter. A partir de esta caracterización se espera que emerjan los proyectos de ciencia ciudadana más importantes, las relaciones que puedan existir entre ellos, y las áreas de conocimiento donde más están impactando estos proyectos. El resultado permitirá determinar, además, que organizaciones, e investigadores están impulsando estas iniciativas.

En esta línea, se ha implementado un sistema informático que realiza el análisis de las conversaciones que se producen en Twitter. El sistema extrae grupos de conversación mediante algoritmos de detección de comunidades, analiza la actividad de publicación y los temas para cada comunidad, y realiza una adecuada visualización de los resultados accesible en línea. Se ha partido de un conjunto de datos que recoge todos los tuits publicados en los seis primeros meses del año 2020 en temas relativos a ciencia ciudadana.

1.3. Alcance

Este proyecto se ha basado en métodos computacionales que han sido desarrollados en el campo de investigación del Análisis de Redes Sociales, Social Network Analysis en inglés, (Wasserman & Faust, 1994; Borgatti et al., 2009). En este contexto, se parte de crear una representación en forma de grafo (red) que relacione a los diferentes actores, usuarios de Twitter en nuestro caso, y las diferentes interacciones que se producen entre ellos, menciones o retuits. Estas redes representan las relaciones globales entre todos los participantes. A partir de esta red global se pueden extraer comunidades de actores explotando el hecho que los actores terminan relacionándose por intereses (por

ejemplo, aficionados a la astronomía que reaccionan a los contenidos publicados por otros usuarios). Una vez definidas las comunidades, es posible explorar las temáticas generadas dentro de cada una de ellas analizando el texto de los comentarios. Según Hoppe (2017), podemos distinguir los tres tipos diferentes de enfoques algorítmicos para analizar las comunidades anteriores y se van a utilizar en este TFM: 1) estructuras de redes que incluyen redes de actores (sociales); 2) series temporales derivadas de la actividad en el tiempo, y 3) contenido utilizando minería de textos.

En el ámbito del análisis de redes sociales (SNA por sus siglas en inglés Social Network Analysis), se han estudiado los procesos de difusión de información en los medios de comunicación social (por ejemplo, Twitter) a fin de identificar a los agentes influyentes en esos procesos de difusión, y los grupos que emergen en estas interacciones. (Leskovec, Backström, Kleinberg, 2009; Agarwal et al., 2012). Estos análisis de comunidades son un medio importante para caracterizar los efectos de las actividades de ciencia ciudadana en la perspectiva de este TFM.

En relación a la minería de texto, se han usado técnicas de procesamiento de lenguaje natural (Qi et al., 2020) que permite extraer información estructurada a partir del texto libre. En particular, se ha empleado reconocimiento de entidades (NER por sus siglas en inglés, Named-Entity Recognition) para poder extraer proyectos y usuarios relevantes en las conversaciones entorno a ciencia ciudadana.

Con el objetivo de combinar los dos enfoques anteriores y enriquecer los resultados, se ha utilizado análisis de redes de textos (NTA por sus siglas en inglés, Network Text Analysis) que permite generar redes extraídas a partir de fragmentos o comentarios donde los nodos son palabras o conceptos, y los enlaces representan apariciones de esos términos en el mismo contexto. En el contexto de este TFM, esto se traduce en el análisis de redes de hashtags.

1.4. Estructura del documento

El documento elaborado consta de las siguientes partes:

1. **Introducción**: pretende exponer el objeto de la realización del proyecto, sentando cuáles son sus fundamentos, los objetivos y el alcance de éste.
2. **Metodología**: se describe el proceso realizado, los datos con los que se han trabajado y cuáles han sido las tecnologías empleadas.
3. **Resultados**: se presentan y comentan los resultados obtenidos.
4. **Discusión**: en ella se analizan y se interpretan los resultados.
5. **Conclusiones y trabajo futuro:** es el cierre del proyecto, explicando los argumentos y afirmaciones sobre el trabajo realizado.

2. Metodología

Para la realización del proyecto se ha seguido un proceso iterativo con dos conjuntos de datos diferentes.

El primero se trata de una extracción semanal de todos los tuits relacionados con ciencia ciudadana desde el 02/05/2020 al 09/05/2020, el cual se ha empleado inicialmente para llevar a cabo un análisis exploratorio de los datos. Contiene información sobre 8.675 tuits. Las gráficas incluidas en este apartado se corresponden a dicho dataset, y se muestran a modo de ejemplo.

El segundo fichero es idéntico en cuanto al formato, pero se trata en esta ocasión de una extracción semestral desde el 01/01/2020 al 30/06/2020, y sobre él se han aplicado los mismos pasos explicado con anterioridad para poder realizar el análisis final. Se comprende de 236.866 tuits. El procesamiento de este fichero dará lugar a los resultados finales, que se expondrán en el tercer apartado de este documento.

El proceso definido puede dividirse en las siguientes tareas:

1. Recopilación de datos: consiste en la extracción de los datos necesarios sobre ciencia ciudadana en Twitter.
2. Ingesta de datos: comprensión de los campos e información contenida en los ficheros de trabajo.
3. Preprocesamiento: se realiza el tratamiento inicial de los datos para evitar errores, como descartar varios nulos o incorrectos.
4. Carga y procesamiento de datos: lectura de los datos del fichero y ejecución de los procesos necesarios para la consecución de los objetivos.
5. Presentación de los datos: se facilita la visualización de los resultados obtenidos al usuario.

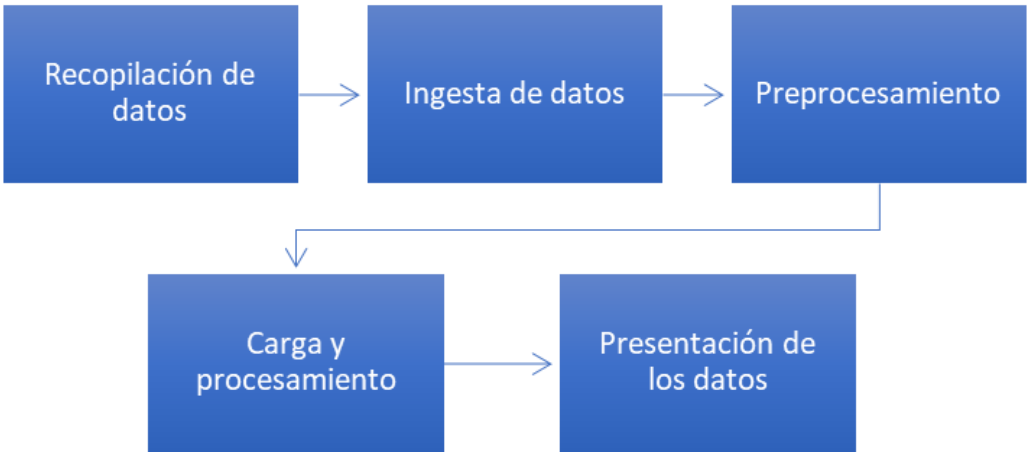

Figura 2-1 Esquema de la metodología

2.1. Recopilación de datos

Estos datos han sido extraídos a través de la herramienta Lynguo (https://www.iic.uam.es/soluciones/entorno-digital/lynguo/) desarrollada por el Instituto de Ingeniería del Conocimiento.

La query (consulta realizada sobre los datos) que se ha empleado contiene las siguientes cláusulas para la búsqueda de tuits relacionados con la ciencia ciudadana:

("citizenscience" OR "citizen science" OR "community science" OR "crowd science" OR "crowd-sourced science" OR "civic science" OR "borgervidenskab" OR "burgerwetenschap" OR "harrastusteadus" OR "kansalaistiede" OR "kansalaisten tied" OR "science citoyenne" OR "sciences citoyennes" OR "sciences participatives" OR "sciences collaboratives" OR "bürgerwissenschaft" OR "scienza dei cittadini" OR "nauka obywatelska" OR "ciencia ciudadana" OR "medborgarforskning" OR "medborgarvetenskap" OR "ciencia ciutadana" OR "herritarren zientzia")

2.2. Ingesta de datos

Ambos conjuntos de datos se han proporcionado mediante ficheros CSV con el mismo formato, en el que cada línea compone el contenido de un tuit, y cuyos campos están delimitados por punto y coma.

Los campos que presentan los ficheros son los siguientes:

- Borrado: indica si el tuit ha sido eliminado o no. Todos los valores proporcionados están a “false”.
- Favorito: indica si el tuit se ha marcado como favorito. Todos los valores proporcionados están a “false”.
- Fecha: muestra el momento de publicación del tuit. El formato definido es YYYY/MM/DD HH:MM:SS. El rango que abarca varía entre la fecha inicial de la extracción y la final. Adicionalmente, es el valor por el que se encuentra ordenado el fichero, de forma que los tuits aparecen desde el de mayor antigüedad al más reciente.
- Fuente: procedencia del tuit. Todos los valores que figuran son “twitter”.
- Texto: contenido del tuit. Es el campo más importante a la hora de extraer información.
- Enlace: enlace de acceso al tuit. No se dispone de la información y todos los valores se encuentra a “null”.

- Marca: indica si el tuit pertenece al ámbito europeo o mundial. Los valores informados son “Citizen science EU” y “Citizen science All World”.
- Usuario: es un valor alfanumérico que identifica al usuario que publica el tuit.
- Oficial: indica si el tuit pertenece a una cuenta oficial (personajes públicos, instituciones, etc.). Todos los valores se encuentran a “false”.
- Opinión: número que valora la opinión sobre el tuit. La escala varía de -100 a 100.
- Categoría: indica los sentimientos del tuit, tales como “alegría”, “amor”, “curiosidad”, “sorpresa”, etc. Muchos de los valores están vacíos.
- Tags: etiquetas que figuran en el tuit. Todos los valores están vacíos. Se extraerán a partir del texto del tuit.
- Impacto: número que determina el impacto del tuit. El rango utilizado es de 0 a 100.

Borrado	Favorito	Fecha	Fuente	Texto	Enlace	Marca	Usuario	Oficial	Opinion	Categoria	Tags	Impacto
false	false	02/05/2020 6:30	twitter	RT: #EarthChallenge2020 is emp	null	Citizen science All World	9a47271916f8dcef8fdf5c7c7bbd3cfe	false	0			56,37
false	false	02/05/2020 6:32	twitter	RT: Also - take part in the newly	null	Citizen science All World	777f9ff7df4e69faf09867bff03d632b	false	0			39,54
false	false	02/05/2020 6:18	twitter	RT: @faliqfahmie Something is fi	null	Citizen science All World	29cc01d0ec55f59f53096e211541d62a	false	-16,67			42,78
false	false	02/05/2020 6:18	twitter	Since there?s no more covid aro	null	Citizen science All World	e73ea4a947a1e3954857737e06c6fa50	false	-33,33			33,87
false	false	02/05/2020 6:19	twitter	RT: @faliqfahmie Something is fi	null	Citizen science All World	f745e7b7588eea0223c65e195709bc55	false	-16,67			36,84
false	false	02/05/2020 6:22	twitter	ok that Azhar Ali kid is no where	null	Citizen science All World	b281933fa28a70b5e97c24dd1f159935	false	0			32,36
false	false	02/05/2020 8:02	twitter	RT: Remember that you can also	null	Citizen science EU	342b5538767b071f64b88bbf381db84c	false	0			39,43
false	false	02/05/2020 8:02	twitter	RT: Remember that you can also	null	Citizen science All World	342b5538767b071f64b88bbf381db84c	false	0			39,43
false	false	02/05/2020 8:02	twitter	RT: @faliqfahmie Something is fi	null	Citizen science All World	832dd95968bcdf0b14ee298a29b7fa00	false	-16,67			34,46
false	false	02/05/2020 7:58	twitter	RT: Please follow @BeesCount -	null	Citizen science All World	f9a82b4be1707203bb87455b539779e5	false	0			4,16
false	false	02/05/2020 7:58	twitter	RT: Can you help us build an urb	null	Citizen science All World	6313f4ef1f3ce4de3336ec7b710f88d6	false	0			26,93
false	false	02/05/2020 7:58	twitter	RT: Can you help us build an urb	null	Citizen science EU	6313f4ef1f3ce4de3336ec7b710f88d6	false	0			26,93

Figura 2-2 Ejemplo cabecera y filas del fichero

2.3. Preprocesamiento de datos

Tras haber comprendido la composición del fichero, se ha realizado un análisis más detallado de los datos que ofrece, cuáles son útiles para el propósito del proyecto, y cómo se debe preparar el fichero para su posterior carga y poder comenzar así con su explotación.

Lo primero que se aprecia es que en el caso de que un tuit cuente con la marca “Citizen science EU”, aparece un registro también duplicado con la marca “Citizen science All World”. Por este motivo, se han eliminado todos los registros con la primera marca mencionada.

El siguiente paso realizado es eliminar aquellos campos que no aportan valor para el objetivo planteado, de forma que permanecerán únicamente los campos “Fecha”, “Texto” y “Usuario”. Posteriormente se detallará el uso que se hace de cada uno de ellos.

También se puede observar que existen algunas filas cuya información está desplazada, no correspondiéndose el valor informado con un valor adecuado del campo, por lo que se han eliminado dichos registros.

Para estos pasos se ha utilizado Microsoft Excel (https://www.microsoft.com/es-es/microsoft-365/excel), que permite visualizar y editar ficheros CSV, filtrarlos para diferenciar los diferentes valores que tiene un campo, eliminar filas o columnas, o eliminar registros duplicados entre muchas otras funcionalidades.

Una vez finalizado, el documento ya se encuentra listo para ser cargado y poder procesar sus datos.

2.4. Carga y procesamiento de los datos

Para la carga y procesamiento de los datos se ha decidido trabajar con Python (https://www.python.org/). Se trata de un lenguaje de programación multiplataforma y multiparadigma, que se caracteriza especialmente por la sencillez y legibilidad de su código. Además, Python está optimizado para trabajar con grandes volúmenes de datos, siendo habitualmente utilizado en proyectos de Big Data y Machine Learning. Por último, hay que destacar que existen multitud de librerías y módulos para Python que proporcionan recursos para el tratamiento de datos, operaciones matemáticas, presentación de los resultados, etc., que van a aportar facilidades y a suponer un importante ahorro de tiempo durante el desarrollo.

El IDE (entorno de desarrollo integrado, por sus siglas en inglés Integrated Development Environment) empleado ha sido PyCharm (https://www.jetbrains.com/es-es/pycharm/) que ofrece ciertas ventajas respecto al uso de la consola. Entre ellas, resaltar el uso de atajos de teclados, guía inteligente para completar código, documentación, detección de errores, o un asistente para la importación de librerías.

Todo el proceso explicado a continuación se ha realizado a partir del dataset de pruebas mencionado con anterioridad.

2.4.1. Carga de datos

Una de las librerías principales utilizadas en el proyecto es Pandas (https://pandas.pydata.org/), que fundamentalmente proporciona y funciones y estructuras de datos para su análisis. El primer uso que se ha realizado de ella ha sido la importación del fichero CSV, que de forma inmediata transforma ya en un DataFrame sobre el que comenzar a trabajar. Esta estructura, habitual en otros lenguajes similares como R, consiste en una colección de columnas con el nombre del campo y el tipo de dato, similar a lo que sería una hoja de cálculo, o una tabla en una base de datos.

2.4.2. Creación de grafos

Con el objetivo de poder detectar las comunidades se ha decidido el empleo de la minería de grafos. Un grafo es una estructura compuesta por un conjunto de vértices V y un conjunto de aristas E en donde los elementos del conjunto de aristas E representan relaciones entre pares de vértices del conjunto V. Dentro de un grafo los vértices pueden representar cualquier objeto, mientras que las aristas representan la relación existente entre esos objetos (Grafo, s.f). Siendo así, el siguiente paso ha sido realizar un procesado de los datos disponibles para transformarlos en una lista de aristas, es decir, representar la información como un grafo.

La primera aproximación será utilizar como vértices las cuentas de los usuarios, y como enlaces las diferentes interacciones que se producen entre ellos. En este caso, a través del usuario publicador del tuit, y del texto de este, se diferencian dos tipos de interacción: un retuit, en el que se publica el tuit de otro usuario, y una cita, en el que simplemente se menciona a uno o varios usuarios.

Para conseguir formar la lista de aristas en el caso de los retuits, se han seguido los siguientes pasos:

1. Se ha filtrado el DataFrame actual para que permanezcan únicamente aquellos registros que comienzan por “RT: @”, puesto que es la cadena que indica que se trata de un retuit.
2. Se ha eliminado la columna Fecha que no tendrá ningún uso para este proceso. Se trabaja entonces con un DataFrame únicamente con dos columnas: “Usuario” y “Texto”.
3. Para la columna Texto, se busca la primera mención realizada, es decir, aquella que comienza por @, y se reemplaza el valor anterior por el nombre de la cuenta.
4. En este punto el usuario está formado por una cadena alfanumérica, mientras que para la cuenta retuiteada se dispone del nombre de esta. Para igualar ambos valores, se ha ejecutado sobre la última el algoritmo MD5, que es el que utiliza Twitter para almacenar el dato.

Para la lista de aristas formada por las menciones, el procedimiento es similar, pero con algunas variaciones:

1. Se descartan aquellos tuits que sean retuits.
2. Se ha eliminado la columna Fecha que no tendrá ningún uso para este proceso.
3. Se busca en la columna texto la primera palabra que comienza por “@”.
4. Se codifica el usuario en MD5.

Una vez que se han creado las dos listas de aristas, se invoca a la función “creategraph”, que se ha desarrollado para que a partir de la lista devuelva un grafo. Para la creación

del grafo se ha recurrido a librería igraph (https://igraph.org/), que aporta todas las facilidades para manejar este tipo de elementos.

La función crea una lista de vértices ordenada a partir de la lista que se ha recibido como parámetro. Se genera un nuevo objeto de clase "Graph", y se le pasan tanto las aristas como los vértices. Se ha otorgado un peso de 1 a todas las aristas, y se ha simplificado el grafo combinando las aristas múltiples sumando sus respectivos pesos.

Inmediatamente después, se llama a la función "plotgraph". El propósito de implementar esta función es mostrar el grafo que recibe como parámetro, por lo que se le ha invocado en ambas ocasiones con el objeto generado en el paso previo. También se le debe pasar el nombre del fichero .png que se quiere generar, y un booleano que va a condicionar si se muestra el nombre de los nodos principales o no.

De cara a visualizar los grafos y para que la representación sea más precisa, los pasos seguidos son los siguientes:

- Se calcula el grado de cada vértice para asignarle un tamaño proporcional. Es decir, cuantos más enlaces tiene, y por lo tanto más importancia, mayor será su tamaño. Para evitar que el dibujo quede desproporcionado, el tamaño se normaliza utilizando la raíz cuadrada de este valor y multiplicándose por un entero.
- El grosor de las aristas se pinta en función de su peso. Cuantas más veces se repita ese enlace, mayor grosor tendrá, hasta un tamaño máximo de diez para evitar ensuciar la figura.
- Si el parámetro "showlabel" recibido a la entrada es "True", entonces se obtienen los diez nodos con mayor peso, se busca el nombre del nodo y se muestra en el dibujo. En ese caso, para no dificultar la comprensión del dibujo, no se altera el grosor de las aristas.
- Se asigna la ubicación y nombre del fichero que se va a generar.
- Se especifica el tamaño y la fuente de las etiquetas de los vértices.
- Se otorga un color a cada vértice en función de su grado.
- Se invoca a la función plot, pasándole el grafo, el nombre, y el estilo de visualización.

A continuación, se muestran un par de ejemplos obtenidos para los RTs y menciones en el dataset de pruebas.

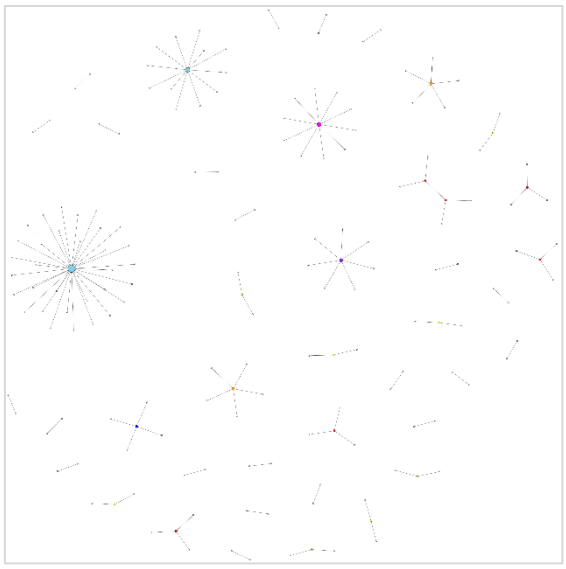

Figura 2-3 Ejemplo de grafo de retuits

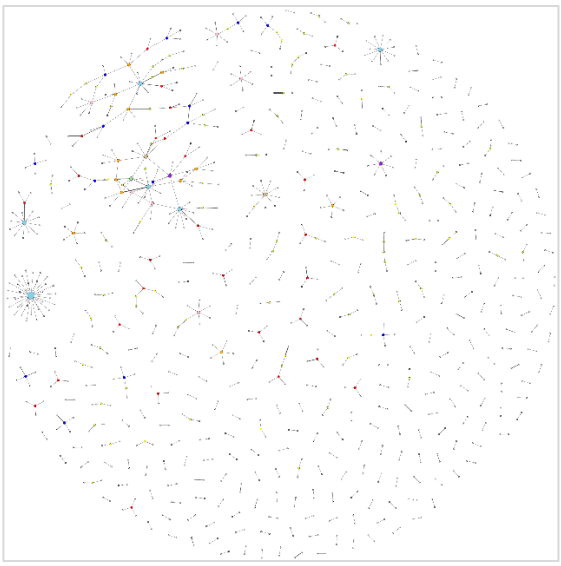

Figura 2-4 Ejemplo de grafo de citas

Debido a que ambos representan interacciones entre usuarios, se han sumado ambas listas de aristas y creado un nuevo grafo combinado a partir de ellas.

Tal y como se explica en el alcance del documento, aplicando el uso de NTA (análisis de redes de texto), se procede a la creación de otro grafo muy significativo, que se obtiene al relacionar los hashtags que aparecen en un mismo tuit. Un hashtag es una etiqueta, que se precede con el carácter "#", empleado para destacar ciertas palabras que son más representativas en el texto. De esta forma se sabe qué hashtags aparecen juntos y por lo tanto estarán relacionados entre sí. Para lograr el objetivo se ha procedido de la siguiente manera:

1. Se extraen los tuits que sean retuits para no repetir información.
2. Se mantiene únicamente la columna "Texto".
3. Por cada hashtag que aparece en el tuit, se añade a la lista de aristas y se relaciona con el resto de los hashtags que aparecen a continuación en el mismo tuit.
4. Se eliminan las aristas que tengan como alguno de sus nodos los hashtags con una sola aparición.
5. Se invoca a la función de crear grafo y se pinta, en esta ocasión con el parámetro "showlabel" a True.

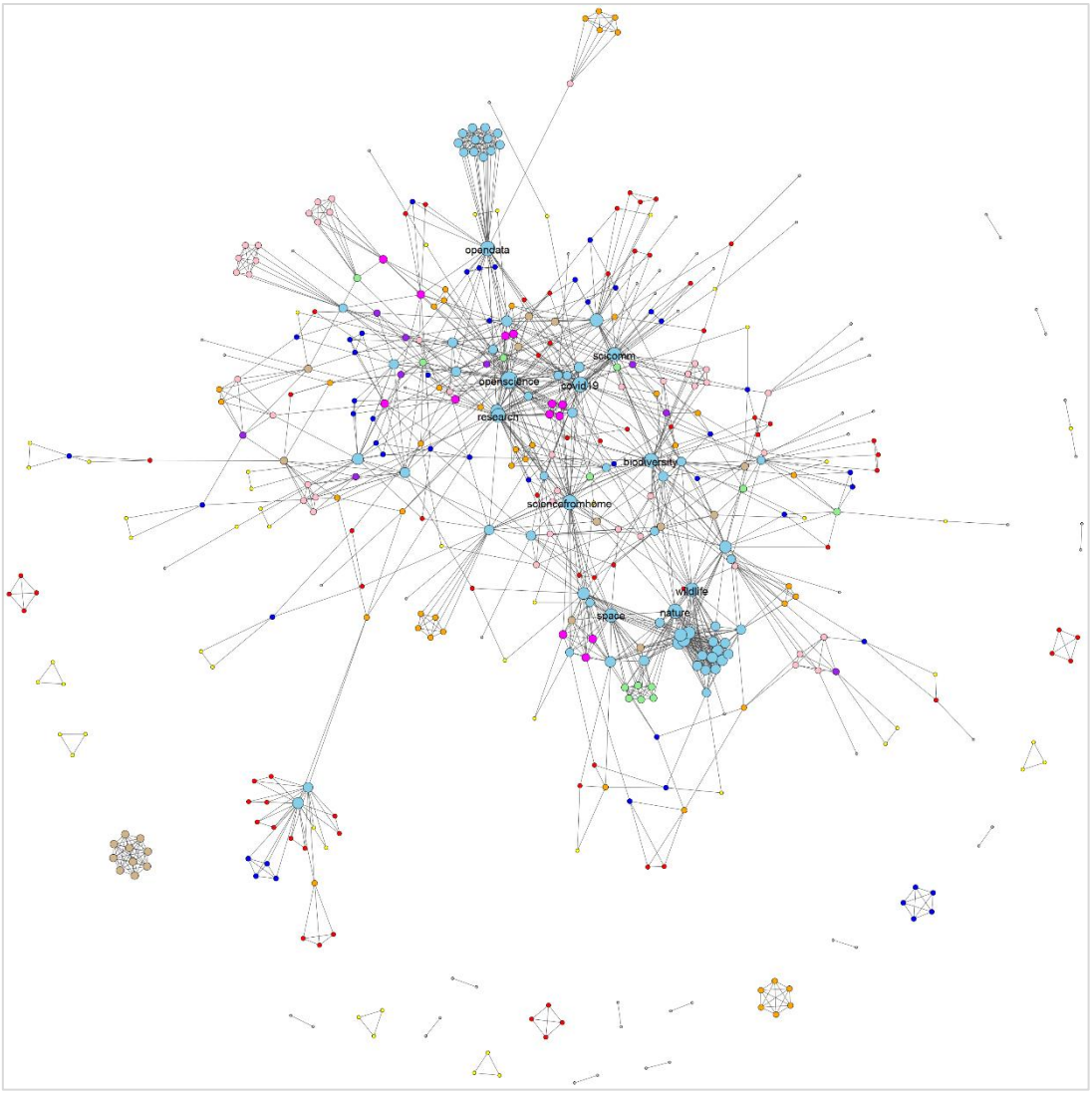

Figura 2-5 Ejemplo de grafo de hashtags relacionados

2.4.3. Minería de textos

Para extraer más información sobre el contenido de los tuits, se han realizado algunos análisis del texto. A continuación, se exponen los gráficos que se han obtenido y los procedimientos utilizados para ello.

Análisis de hashtags:

Por ejemplo, aprovechando las dos listas que se han formado anteriormente para crear los grafos, se ha procedido a extraer aquellos hashtags que más veces se han retuiteado y los que más se han utilizado como hashtags principales.

Para el primero de ellos, partiendo de la lista de tuits que son retuit, se extraen todas las palabras que comiencen por "#". Después se convierten a minúsculas todos para evitar que el mismo hashtag aparezca escrito de diferentes formas, y se eliminan los

hashtags relacionados con ciencia ciudadana , similares a “citizenscience”, por ser este el criterio de búsqueda principal en la extracción de datos. Por último, se transforman a una lista que contiene los hashtags y el número de veces que aparece en total.

Una vez que se ha obtenido la información, se ha definido la función “plotbarchart” para representarla en un gráfico de barras. Esta función recibe como parámetro: el número de hashtags que se desean mostrar, la lista con el número de apariciones de cada hashtag, la lista de hashtags, el título del gráfico, la etiqueta del eje X, la etiqueta del eje Y, y el nombre del fichero.

La función se apoya en la librería Matplotlib de Python, que es una librería gráfica para representación de datos, y en Seaborn, que es una extensión de ésta que aporta mayor funcionalidad para la visualización. Además de establecer la información con los parámetros de entrada, se ha seleccionado el tamaño de la figura, tamaño y estilo de la fuente para el título y etiquetas, ángulo de las etiquetas, la escala de los datos y el color de las barras.

Para la creación de la gráfica de hashtags más utilizados, se ha hecho uso del DataFrame en el que ya se habían eliminado los retuits. El procedimiento seguido a continuación es el mismo que en el caso anterior.

Por último, se ha realizado también el análisis global de los hashtags más utilizados contando todos los tuits, que representará la suma de los dos gráficos previos.

Análisis de términos más frecuentemente usados:

Otro de los gráficos que se ha realizado es el de las palabras que más aparecen en el texto. Para ello se ha hecho uso de una librería de Python llamada Stanza, que contiene numerosas herramientas para el procesamiento de lenguaje natural. En concreto, se ha centrado en el uso de NER, que como se menciona en la introducción, es el reconocimiento de entidades nombradas. Se trata de una tarea para categorizar cada una de las entidades del texto, tanto a nivel del tipo de palabra (si es un verbo, nombre, adjetivo, artículo…) como a nivel de ciertas categorías predefinidas (personas, organizaciones, lugares…).

Al comenzar con el uso de esta librería, el primer paso es construir un Pipeline, en el que se debe indicar el idioma con el que se va a trabajar, y los procesadores que se utilizarán. En esta ocasión, el idioma indicado es el inglés, pues es el mayoritario en el conjunto de datos, y ha sido necesario el uso de tres procesadores: TokenizeProcessor, que divide el texto en tokens y frases; POSProcessor, que permite el acceso a ciertas propiedades de las palabras, como el tipo de palabra, y por último NERProcessor, que como se ha comentado previamente, va a permitir categorizar las palabras.

De forma que, a nivel de código, se debe descargar inicialmente el modelo en inglés y los procesadores indicados, y posteriormente se crea el Pipeline para cargarlos. Después se le pasa al Pipeline el texto que se desea analizar y se obtiene un objeto de tipo

Document. Para ello se han concatenado todos los tuits, añadiendo dos retornos de carro para separarlos. El objeto Document se divide en frases, tokens y palabras. Por lo tanto, una vez recorrida cada frase, utilizando el procesador POS, se han obtenido únicamente las palabras que sean adjetivos, verbos o nombres, por considerarse de mayor importancia semántica. Se convierten a minúscula, se eliminan aquellas relacionadas con "ciencia ciudadana", y se añaden a una lista de forma única con el número de ocurrencias que tiene cada una de ellas.

Como ejemplo de los gráficos realizados, se muestra a continuación una de las gráficas obtenidas.

Figura 2-6 Ejemplo de palabras con más ocurrencias

Análisis de evolución temporal:

Para terminar con el análisis genérico del texto, aprovechando que se dispone de la columna "Fecha" con el momento de publicación de cada tuit, se ha decidido realizar un estudio de la evolución temporal de los hashtags retuiteados y principales.

El primer paso es decidir la escala en la que se representará la información. Debido a que en el dataset de pruebas se dispone de la información de una semana, se agruparán los hashtags de forma diaria. Para ello se extraen todos los valores del campo "Fecha", se eliminan los valores duplicados para quedarnos con los valores de forma única, y se ordenan para que estén disponibles del más antiguo al más actual.

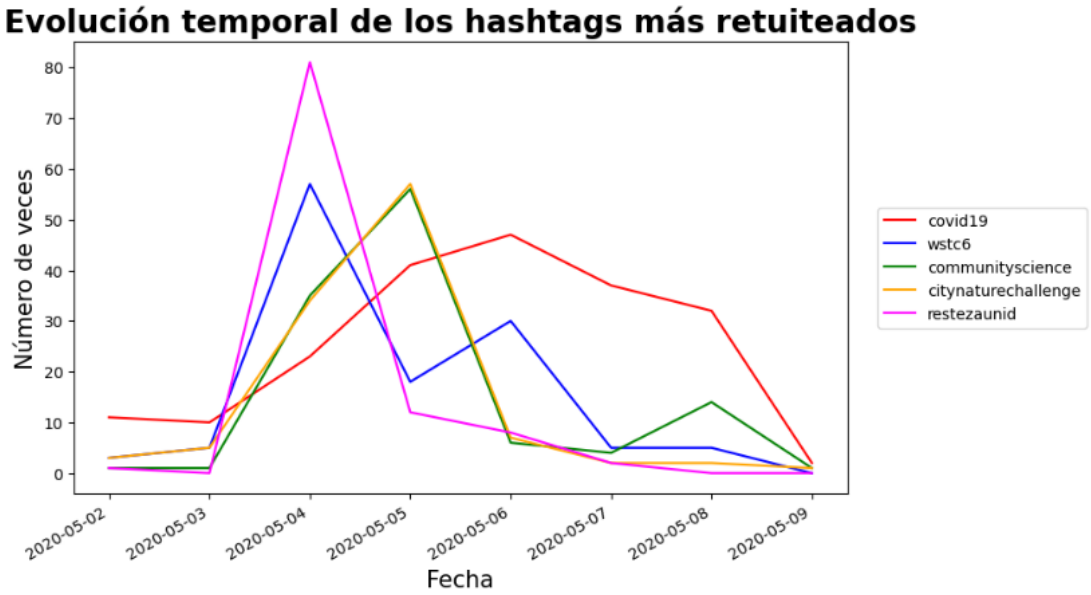

Figura 2-7 Ejemplo de evolución temporal hashtags retuiteados

A continuación, se ha invocado a la función "plottemporalserie", definida para pintar la gráfica temporal. Esta función recibe como parámetros: lista de elementos de tiempo, DataFrame a analizar, lista de hashtags, nombre del fichero y título. A partir de aquí, se recorren los cinco primeros hashtags (que son los de mayor aparición), y se divide el DataFrame por cada uno de los elementos temporales para obtener los tuits de cada día. Después se recorren para contar el número de veces que aparece el hashtag que se está analizando.

Para mostrar la información, se emplea la librería Matplotlib, pero en esta ocasión utilizando la función "plot_date". Mediante los parámetros de entrada es posible configurar la leyenda que se quiere indicar, los colores de las gráficas, y el valor de ambos ejes. También, de forma independiente, se define el título, el nombre de ambos ejes, el estilo y tamaño de fuente, el ángulo de las etiquetas, y la ubicación de la leyenda.

2.4.4. Detección de comunidades

El objetivo principal del proyecto es la detección de las comunidades de ciencia ciudadana en Twitter. Para lograrlo, se ha decidido trabajar sobre el grafo de hashtags ya que por cada tuit se crean más enlaces que respecto al grafo de interacciones, y por lo tanto aportará más información a partir de los mismos datos.

Para comenzar, se han extraído los clústeres a partir del grafo. Un clúster es un conjunto de objetos cuyos miembros se asignan a un mismo grupo por tener características similares. En este caso, porque existen relaciones entre los hashtags. Se pueden obtener fácilmente a través de una función de la librería igraph que devuelve los clústeres de un grafo.

Se han eliminado aquellos clústeres que cuentan con menos de cuatro componentes por no considerarse de suficiente relevancia.

Posteriormente se invoca a la función “plotcommunities”, en la que se obtienen las comunidades y se pintan en el grafo. En este punto, el grafo está compuesto, por un lado, por una componente conexa principal, que es aquella en la que existe un camino de aristas por el que se puede llegar entre cualquier par de vértices. Por el otro, existen diferentes componentes aislados más pequeños, que constituyen ya por sí mismos una comunidad. Para la componente conexa se han ejecutado algunos algoritmos de detección de comunidades que proporciona la librería graph.

Inicialmente se ha probado con el método **Edge betweenness** (Dubitzky, Wolkenhauer, Cho, Yokota, 2013), el cual asigna a cada arista una puntuación en función del número de caminos cortos que existen entre vértices y que contengan dicha arista. De esta forma, se entiende que cuanto mayor es la puntuación, más probabilidad existe de que ambos nodos pertenezcan a diferentes comunidades. De cara a la representación gráfica, los criterios seguidos han sido los mismos que con los grafos anteriores, salvo que, en esta ocasión en vez de asignar un color en función del grado del nodo, se han pintado del mismo color aquellos nodos que pertenecen a una misma comunidad, y se muestra siempre su hashtag más representativo.

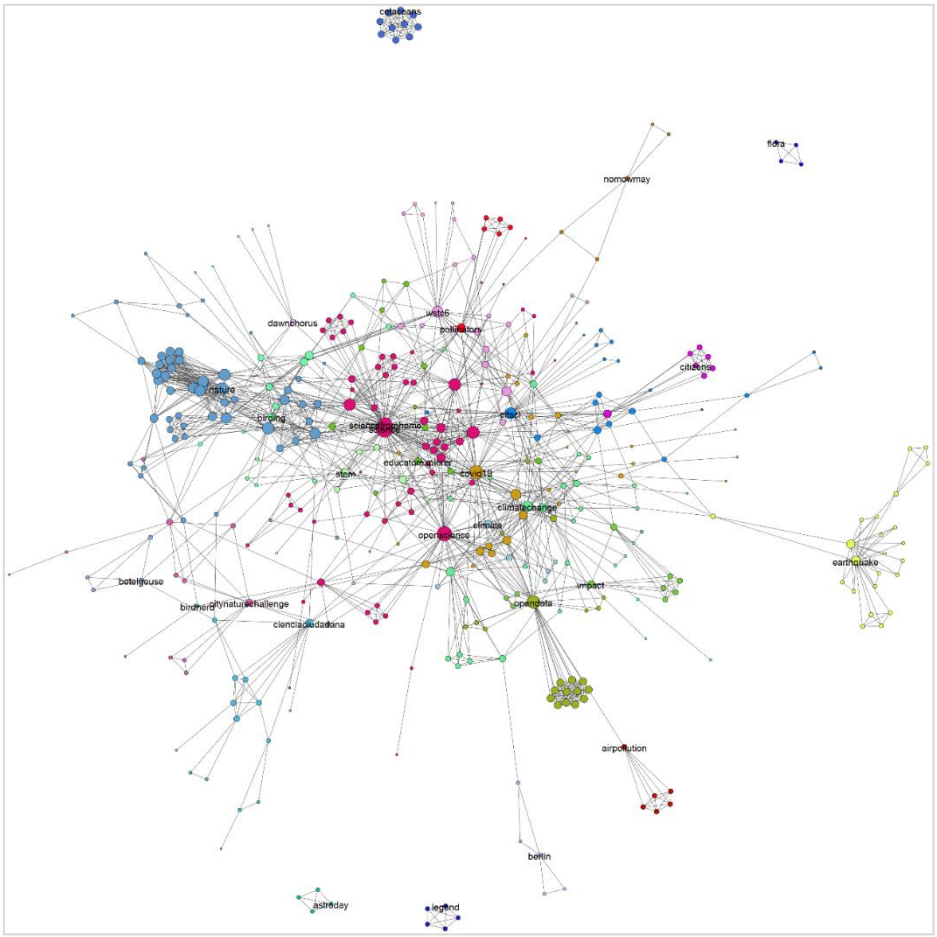

Figura 2-8 Ejemplo de comunidades detectadas mediante Edge betweenness

Para finalizar, la función ordena las comunidades de mayor a menor número de componentes, elimina aquellos componentes con menos de cuatro elementos, y devuelve la lista como resultado. También se exportan los datos a un fichero en el siguiente formato:

Comunidad 1:

Hashtag 1

Hashtag 2

Hashtag 3

Comunidad 2:

Hashtag 1

Hashtag 2

…

Además, el mismo proceso se ha realizado con otros dos algoritmos muy populares en la detección de comunidades: Walktrap (Pons & Latapy, 2005) y Label Propagation (Zhu, Ghahramani, 2002).

Walktrap es un algoritmo ideado por Pons y Latapy, que realiza un agrupamiento jerárquico buscando aquellos vértices comunicados por caminos cortos, a través de una serie de caminos aleatorios. La probabilidad de conectar desde un nodo a otro es proporcional al peso de los enlaces, y será mayor si ambos nodos están fuertemente relacionados. Este algoritmo tiene la restricción de que sólo es posible asignar un nodo a una única comunidad.

En igraph, el algoritmo permite introducir opcionalmente dos parámetros de entrada: el peso de las aristas, y los pasos en la longitud de los caminos aleatorios que se llevarán a cabo. La primera prueba se ha realizado con el peso, y se han fijado a cuatro los pasos, ya que generalmente se utilizan valores entre tres y cinco.

Adicionalmente, en esta ocasión, al invocar a la función de visualización del gráfico, se ha indicado la utilización del parámetro “mark_groups” a verdadero, lo que destaca mediante formas coloreadas cada una de las comunidades.

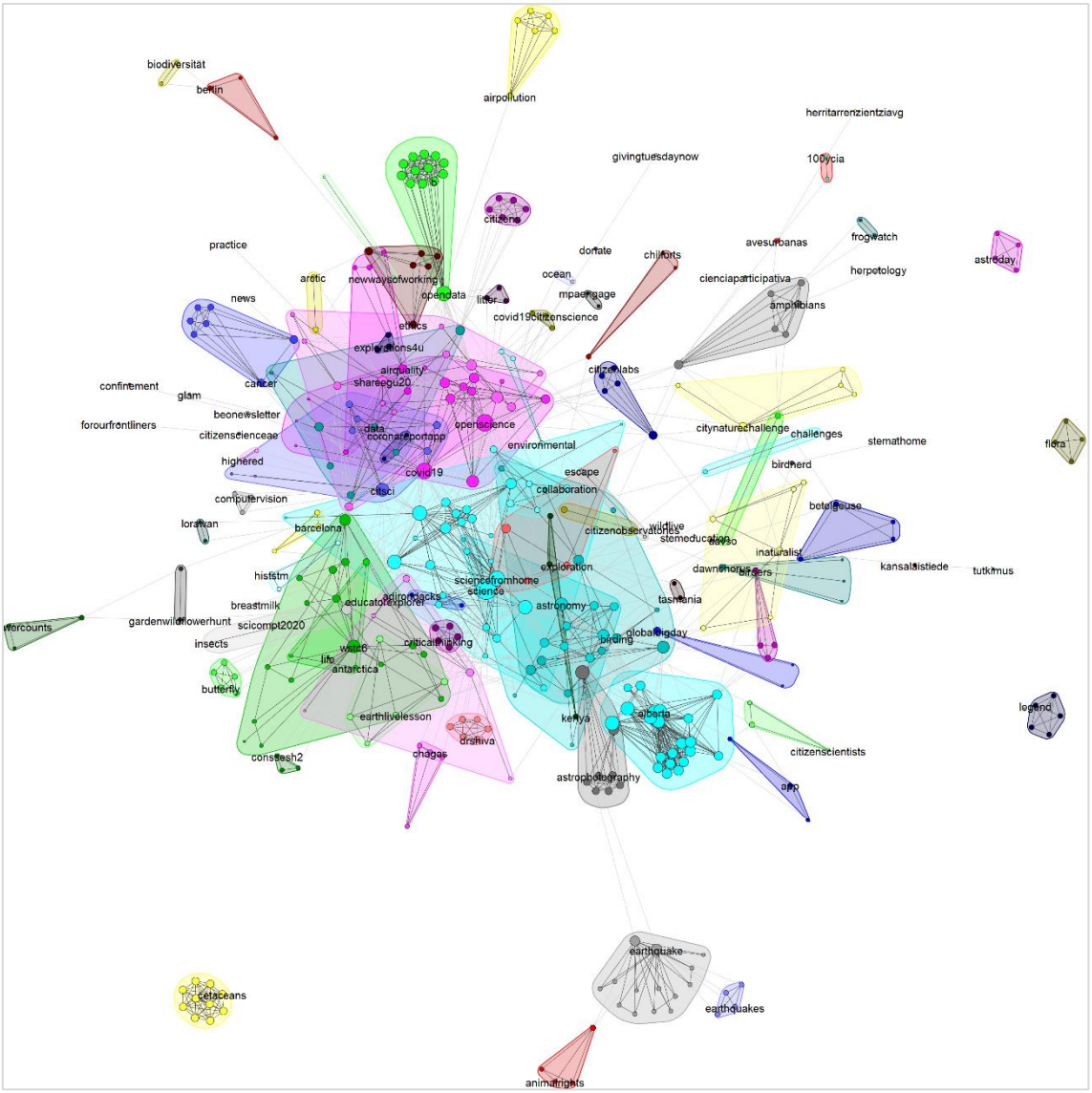

Figura 2-9 Ejemplo de comunidades detectadas mediante Walktrap con pesos

Label propagation, a diferencia de la mayoría de algoritmos cuyo objetivo se basa en la optimización de alguna métrica, se guía utilizando la estructura del grafo. Al inicio, se asigna a cada nodo una etiqueta aleatoria, que se va propagando a lo largo de una serie de iteraciones. Por cada una de ellas, la etiqueta de cada nodo se actualiza con el valor más frecuente de las etiquetas de sus vecinos. El proceso finaliza cuando todas las etiquetas asignadas son mayoritarias entre sus vecinos. Las comunidades se forman entonces por aquellos nodos cuya etiqueta es la misma.

Ambos algoritmos se han ejecutado también sin tener en cuenta el peso de las aristas, lo cual restará importancia al hecho de que ciertos hashtags se repitan con mucha frecuencia y limita la búsqueda únicamente a que estén relacionados.

Mientras que en Label Propagation apenas se ve afectado el resultado, en Walktrap, al omitir el peso, se establecen prácticamente el doble de comunidades.

2.4.5. Información sobre las comunidades

Una vez que se dispone de la lista de las comunidades y sus respectivos hashtags, el objetivo es extraer las organizaciones, personas y proyectos que aparecen en las comunidades obtenidas.

Para lograrlo, se ha enriquecido el fichero de datos, asignando por cada tuit las comunidades a las que pertenece. Para ello se recorre el DataFrame, se comparan los hashtags que aparecen en el texto del tuit con los que contienen las comunidades, y en el caso de que coincidan, se añaden al valor a la columna "Comunidades" de dicho tuit.

Después se recorren los tuits de cada comunidad, y se extraen todas las entidades utilizando el módulo NER de Stanza. Inicialmente se filtran todas las palabras relacionadas con "citizen science", y posteriormente se seleccionan aquellas cuyo tipo sea ORG (Organización) y PERSON (Persona).

Además, de cara a obtener los proyectos se han extraído aquellas entidades que contienen la palabra "project" o "program".

Por último, se obtienen los objetos únicos de cada una de las listas creadas para evitar entidades duplicadas.

El resultado final se exporta a un fichero, del cual se muestran a continuación un par de comunidades a modo de ejemplo:

Comunidad 9:

***Organizaciones**: [' '#EducatorExplorer' '#NBCTStrong' 'GEMS World Academy-Chicago' 'Ginkgo' 'Globe' 'Life Science' 'NBCT' 'NGSS' 'NatGeoEducation' 'OxZooDept' 'R&D' 'Rudee Maniacs' 'Twitterchat' 'the #StayHomeBiodiversityChallenge' 'the Directed Energy Research Centre' 'the Technology Innovation Institute']*

***Personas**: ['Amy Downs' 'Beth' 'DebrisTracker' 'Felix Vega' 'Fiona Jones' 'Jill' 'LillygolS' 'Mary' 'MattHHolden' 'Peg Keiner' 'Rosetta']*

***Proyectos**: ['Electromagnetic Programs']*

Comunidad 13:

***Organizaciones**: ['@PoMScheme' 'BBC' 'DaveGoulson and @The_Buzz_Club' 'FIT' 'Facebook']*

***Personas**: ['Claire']*

***Proyectos**: ['the Lost Ladybug Project']*

2.5. Presentación de los datos

Para la presentación de los datos obtenidos datos se ha decidido la realización de un sitio web que permita la visualización y navegación de los usuarios por los resultados de manera online.

3. Resultados

Una vez que se dispone de la metodología, se procede a repetir el proceso completo, esta vez utilizando el dataset final que contiene los tuits relativos a ciencia ciudadana de los seis primeros meses del año 2020.

3.1. Resumen de los datos procesados

Tras haber realizado el procesamiento de los datos, se muestra una ficha a modo de resumen para la comprensión del volumen de la conversación global.

	Número
Tuits	175.857
Participantes	74.303
Usuarios únicos mencionados	11.732
Hashtags	493.016
Hashtags únicos	18.581

Tabla 3-1 Resumen de los datos procesados

3.2. Creación de grafos

A continuación, se exponen los grafos realizados a partir de las interacciones de usuarios. En esta ocasión, debido a la limitación encontrada para mostrar los grafos completos en igraph, puesto que a partir de cierto número de elementos el gráfico se distorsiona, se ha exportado la lista de aristas y utilizado un software libre para la visualización de grafos llamado Gephi (https://gephi.org/).

Para simplificar la lectura de estos grafos, y poder leer la información más importante que contienen, se ha filtrado y seleccionado su componente gigante (que es la componente conexa de mayor tamaño) para representarlos. Se adjuntan en el Anexo II un par de ejemplos del grafo completo por si se desean visualizar.

En el grafo de retuits se muestran todos los elementos que forman la componente gigante, mientras que para el grafo de citas y el combinado se han filtrado aquellos nodos cuyo grado es menor a ocho y menor a diez respectivamente.

Además, se adjunta en el Anexo III la distribución de grados para cada uno de ellos.

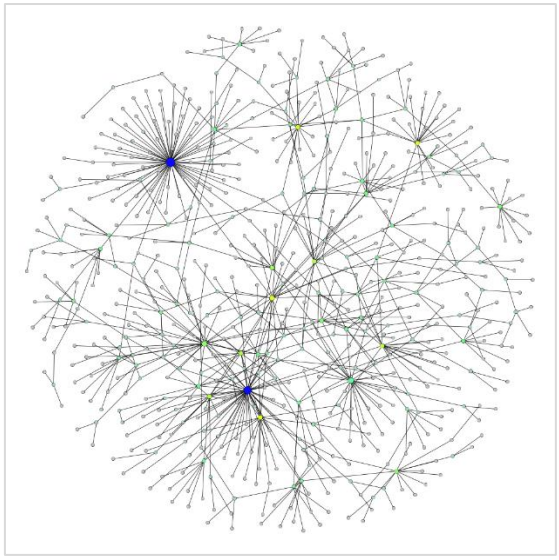

Figura 3-1 Componente gigante del grafo de retuits

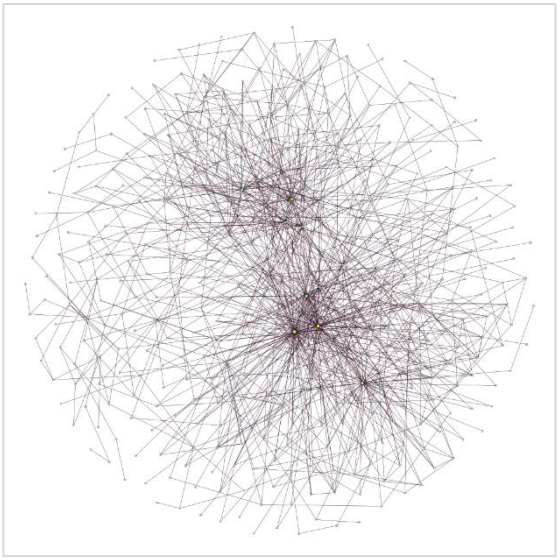

Figura 3-2 Componente gigante del grafo de citas

Combinando ambos grafos mediante la suma de sus aristas, se ha obtenido un nuevo grafo que representa el total de interacciones entre todos los usuarios.

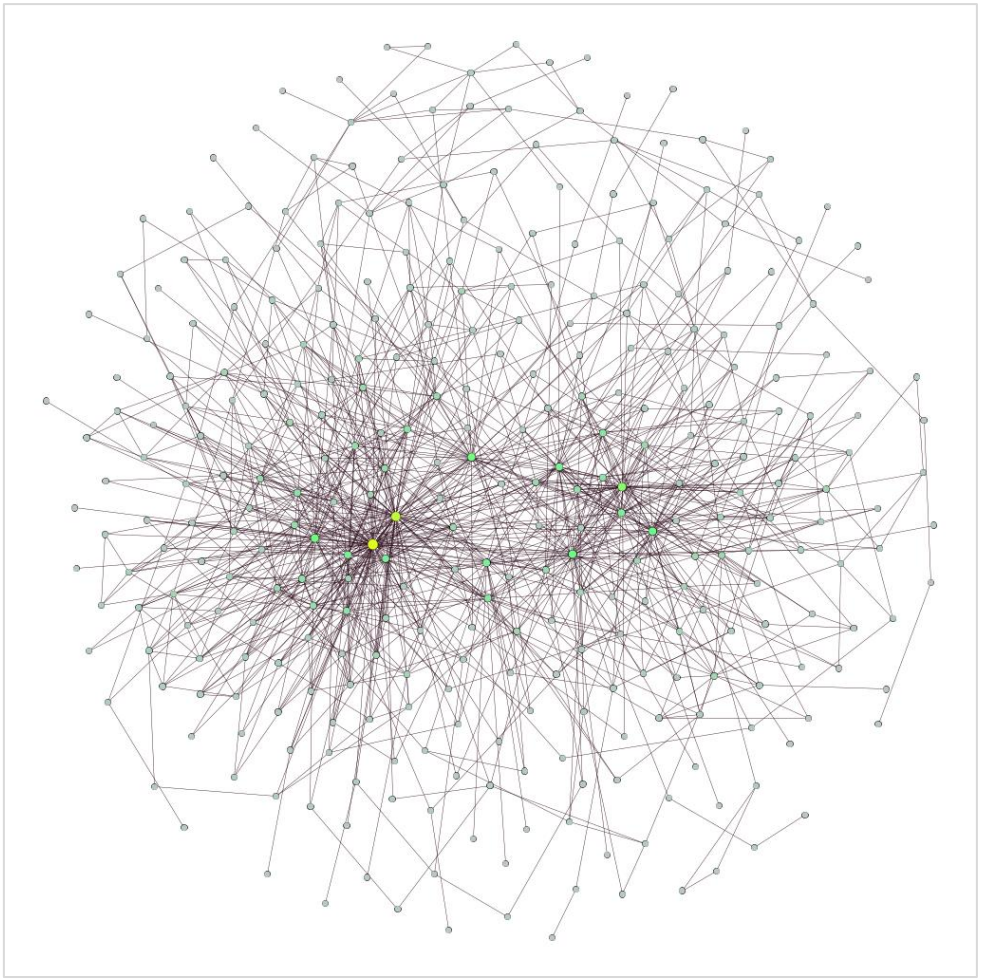

Figura 3-3 Componente gigante grafo combinado

Por último, se visualiza el grafo que relaciona aquellos hashtags que aparecen en un mismo tuit, utilizando una muestra representativa que supone más de la mitad de los elementos totales del grafo.

Se complementan los grafos obtenidos con la Tabla 3-1, en la que se muestra el número de vértices y aristas por los que está compuesto cada uno de ellos.

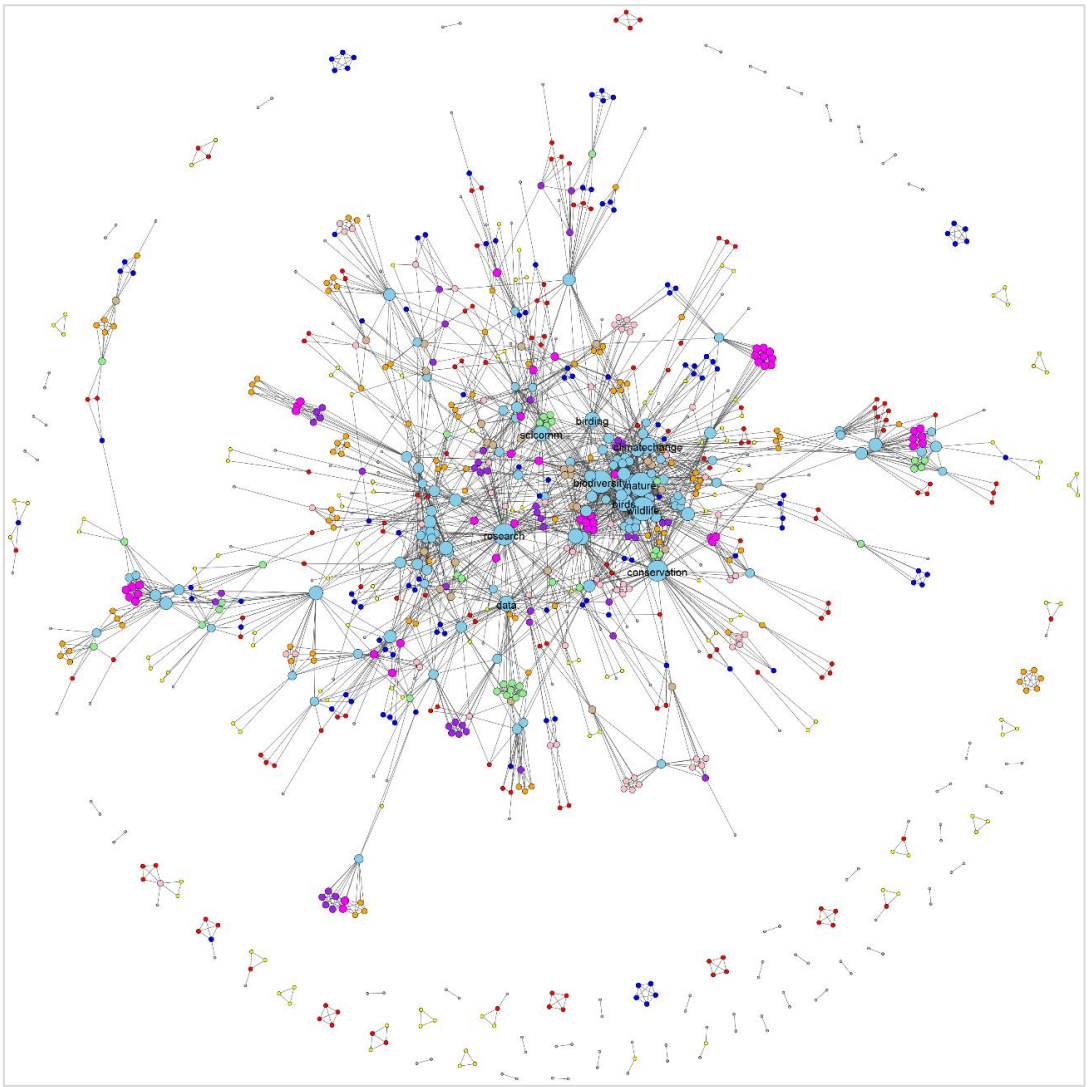

Figura 3-4 Grafo hashtags

Grafo	*Nº de vértices*	*Nº de aristas*
Retuits	2686	2145
Citas	21524	19327
Combinado	22799	21299
Hashtags	6160	426552

Tabla 3-2 Información sobre grafos

3.3. Minería de textos

Se muestran las siguientes gráficas obtenidas mediante técnicas de minería de texto.

En el caso de los hashtags que más veces se retuitean, se han analizado 102.676 tuits, dando lugar a un total de 11.660 hashtags diferentes.

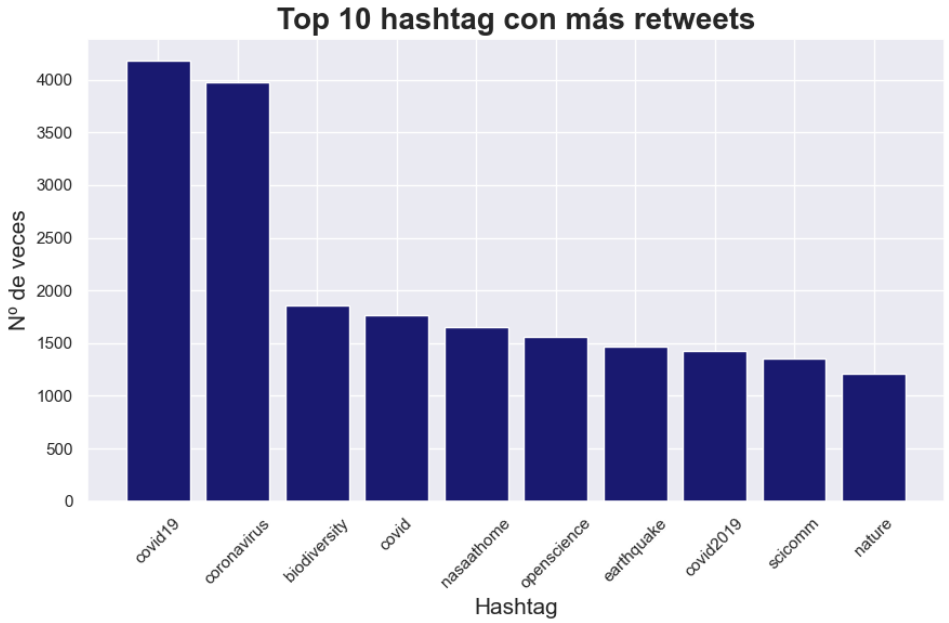

Figura 3-5 Hashtags con más retuits

Para los hashtags principales, se han procesado 73.181 tuits y se obtiene un total de 18.216 hashtags distintos.

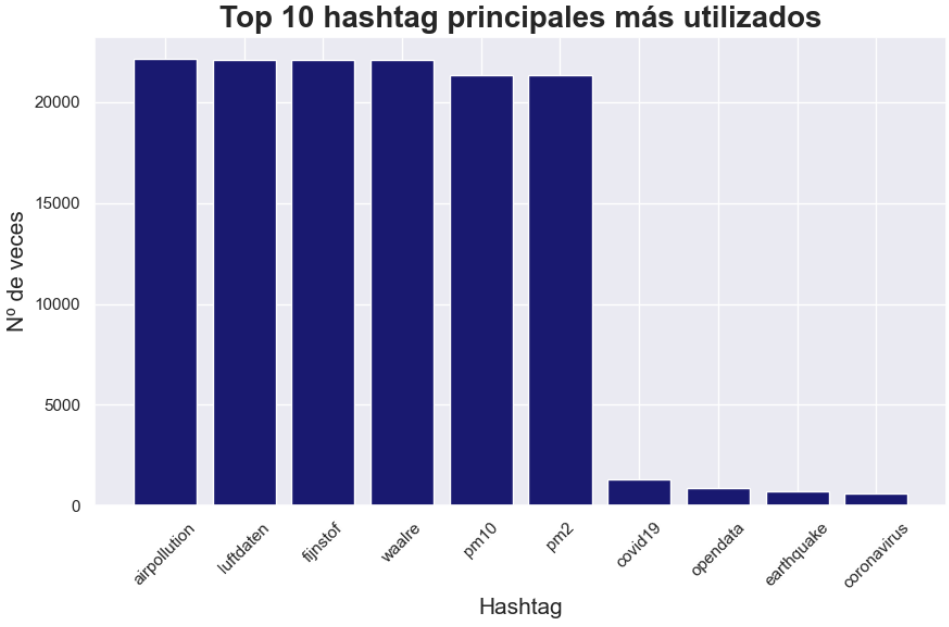

Figura 3-6 Hashtags principales más utilizados

Debido a que existen seis hashtags con un número muy superior al resto que se publican conjuntamente en el mismo tuit de forma repetida, se incluye también la gráfica para los diez valores siguientes con el objetivo de presentar una información más completa sobre el contexto de la información.

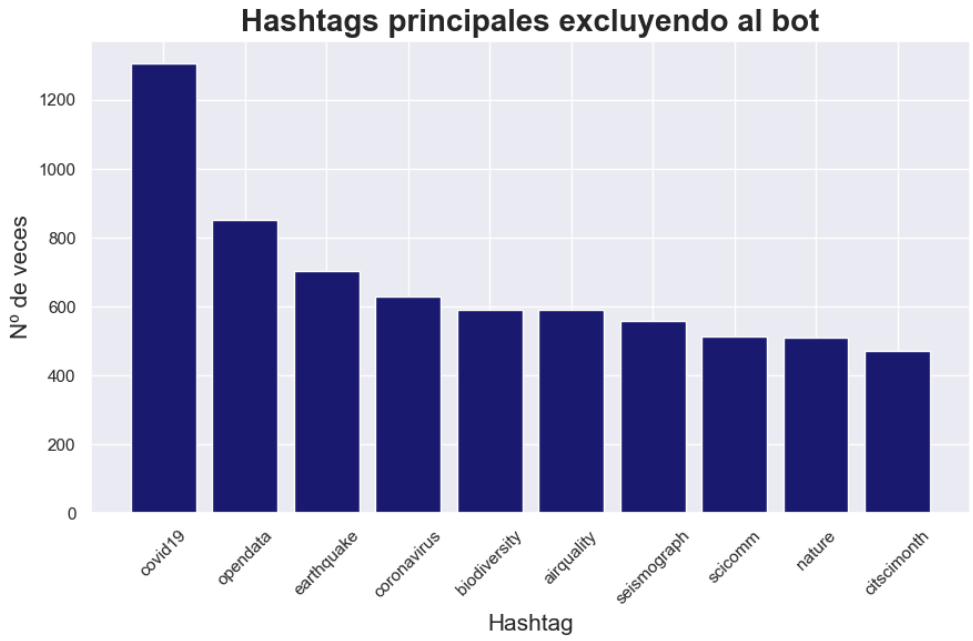

Figura 3-7 Top hashtags excluyendo aquellos utilizados por el bot

Los hashtags totales, tanto de tuits principales como de retuits, se han obtenido tras buscar, en los 175.857 tuits que componen el dataset completo, todas las palabras que comienzan por "#", que tras eliminar los valores repetidos han sido 18.581 hashtags únicos.

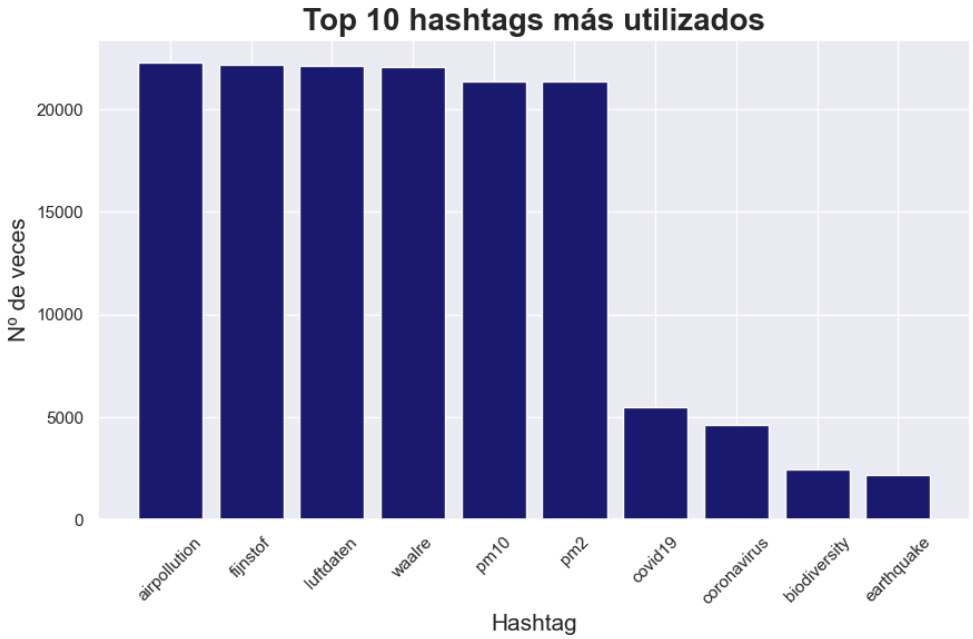

Figura 3-8 Hashtags totales con más ocurrencias

En el proceso de detección de los términos más frecuentes en el texto se han analizado 5.570.348 palabras.

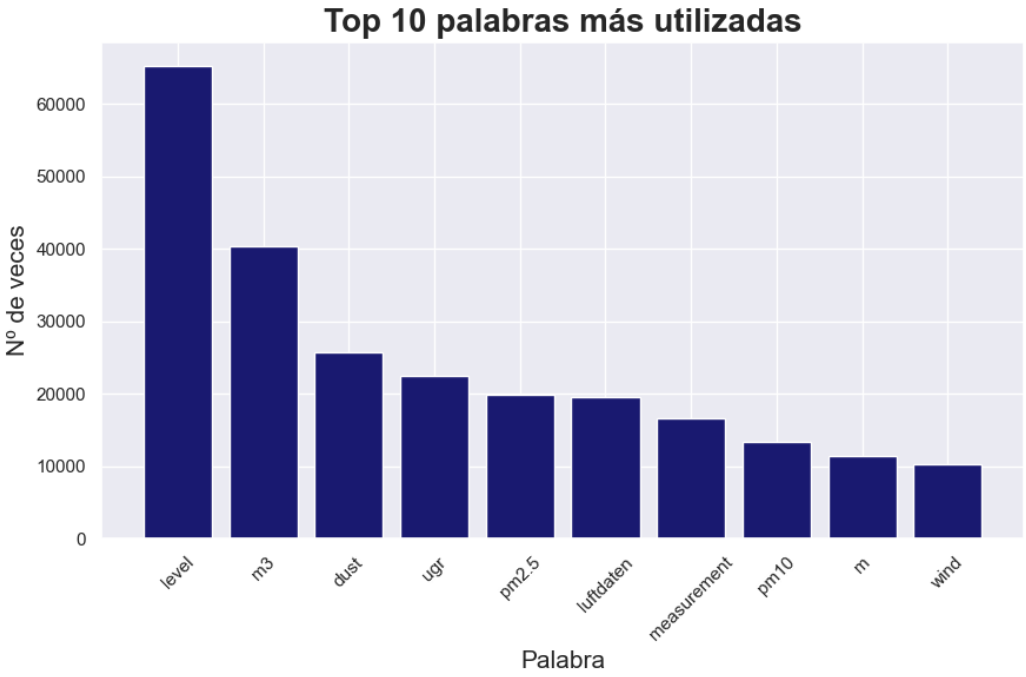

Figura 3-9 Términos con más ocurrencias

También se ha representado la evolución temporal de los cinco hashtags más retuiteados y los más utilizados a lo largo de la primera mitad del año.

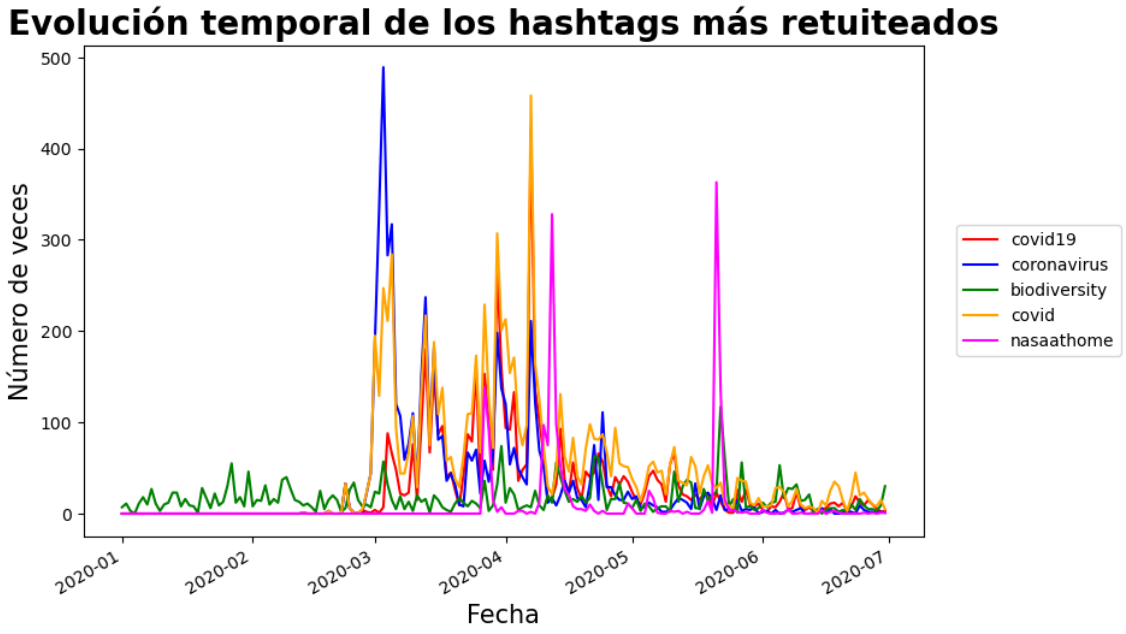

Figura 3-10 Evolución temporal de los hashtags más retuiteados

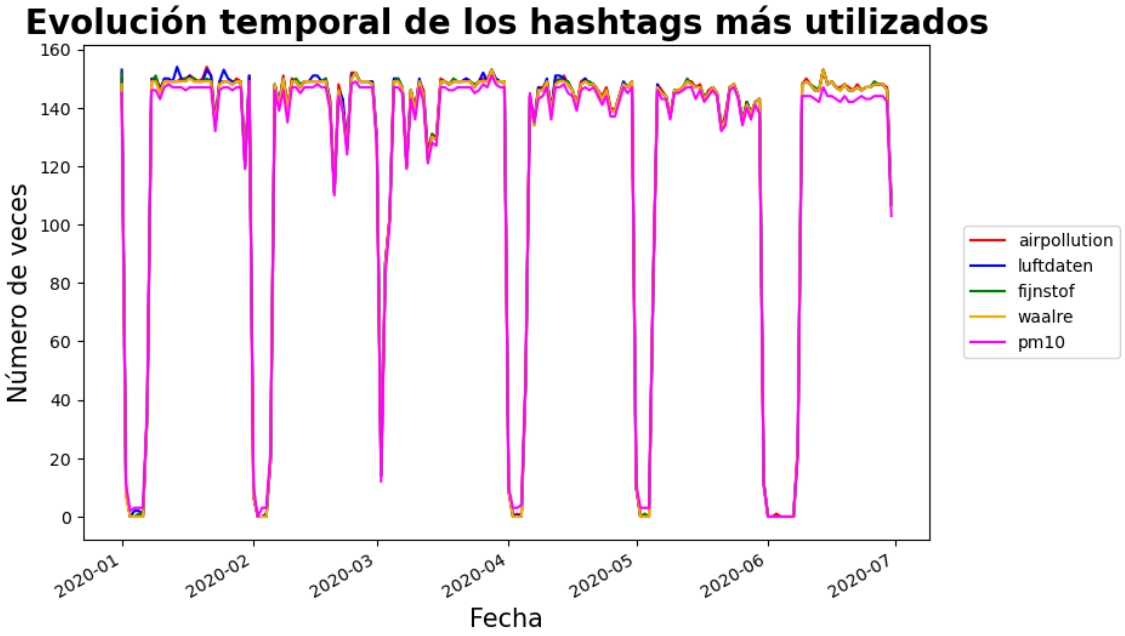

Figura 3-11 Evolución temporal de los hashtags más utilizados

De la misma forma que se ha realizado la obtención de los diez hashtags siguientes en la Figura 3-7, se muestra ahora su evolución en el tiempo.

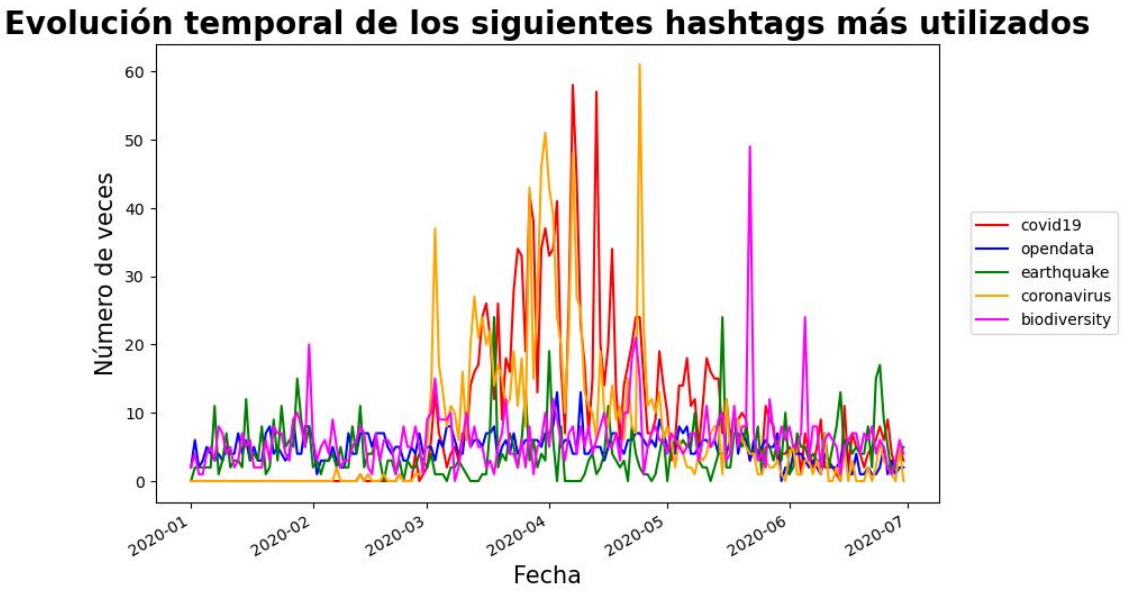

Figura 3-12 Evolución temporal de los siguientes hashtags más retuiteados

3.4. Detección de comunidades

Las siguientes gráficas que se muestran a continuación son el resultado de la ejecución de los diferentes algoritmos de detección de comunidades aplicados al set de datos final. Adicionalmente, en el repositorio indicado en el Anexo I se han subido los ficheros con las comunidades obtenidas y la lista de hashtags por la que están formadas. El nombre de los ficheros es "grafoAlgoritmoDeGeneración.txt".

Como complemento a esta información, se incluye la Tabla 3-3, mostrando el número de comunidades detectadas en cada caso y el número medio de hashtags por comunidad.

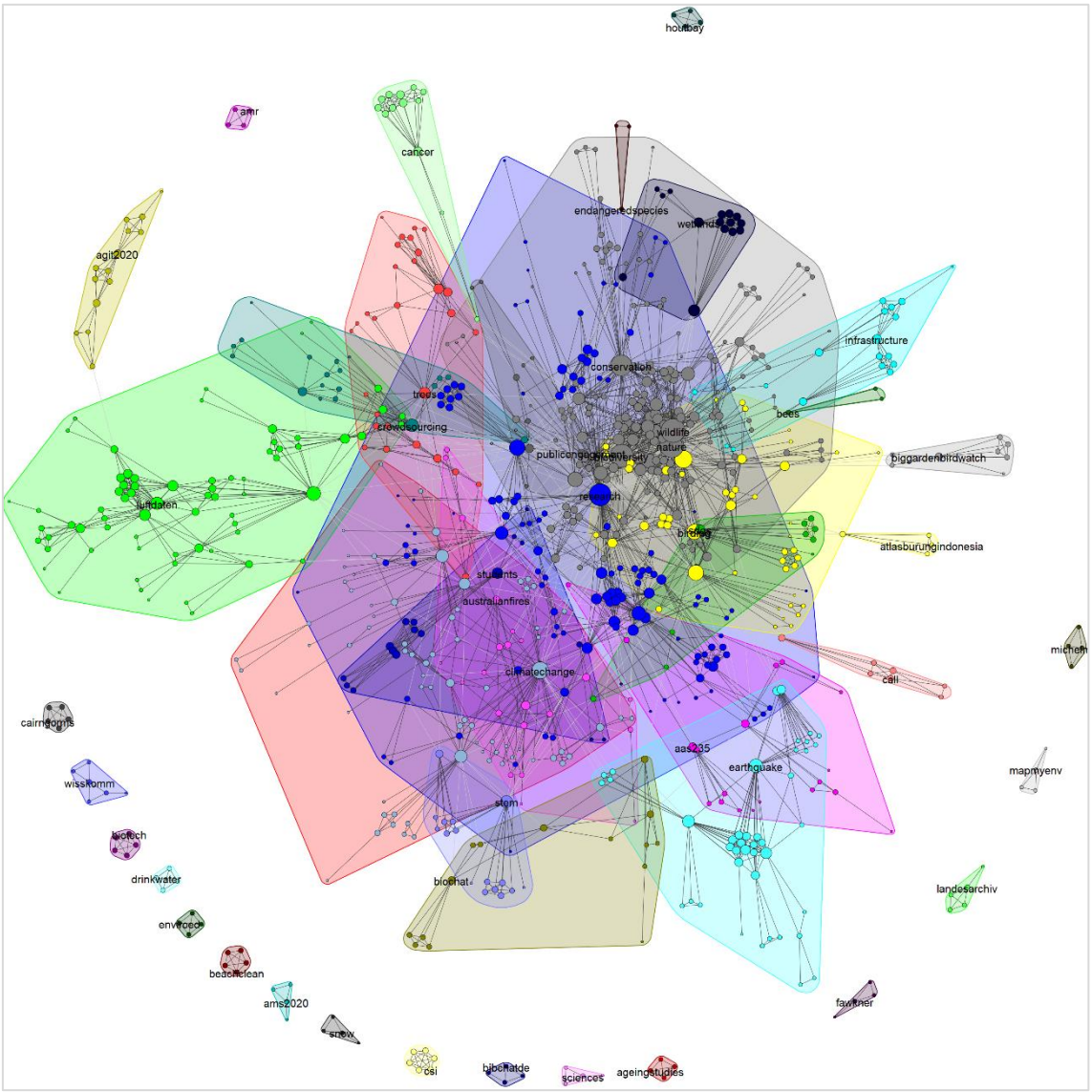

Figura 3-13 Comunidades detectadas mediante Edge betweenness

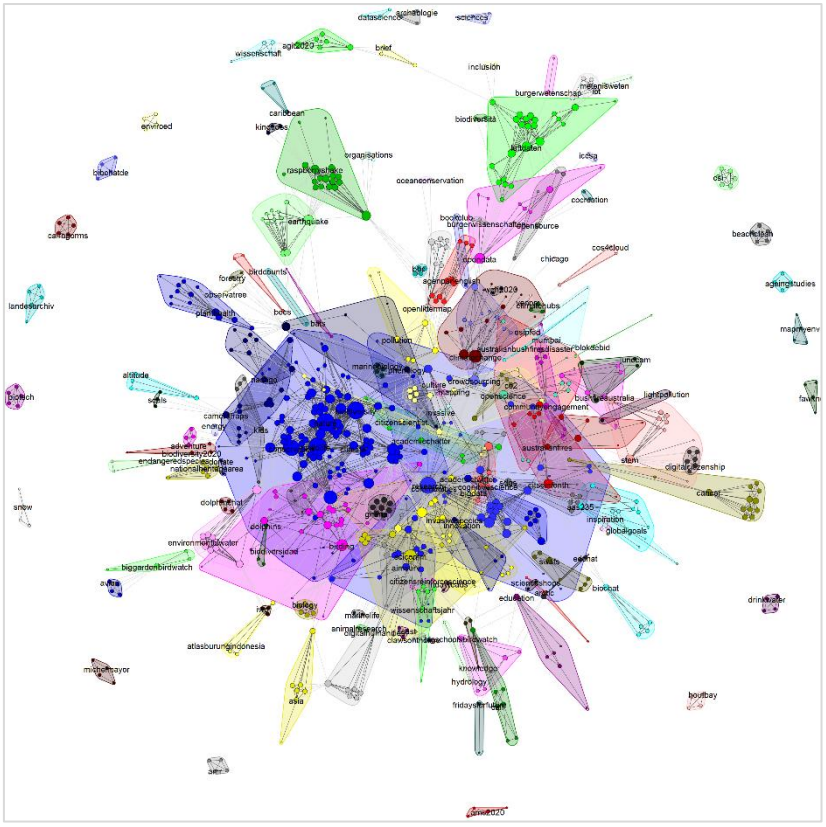

Figura 3-14 Comunidades detectadas mediante Walktrap con pesos

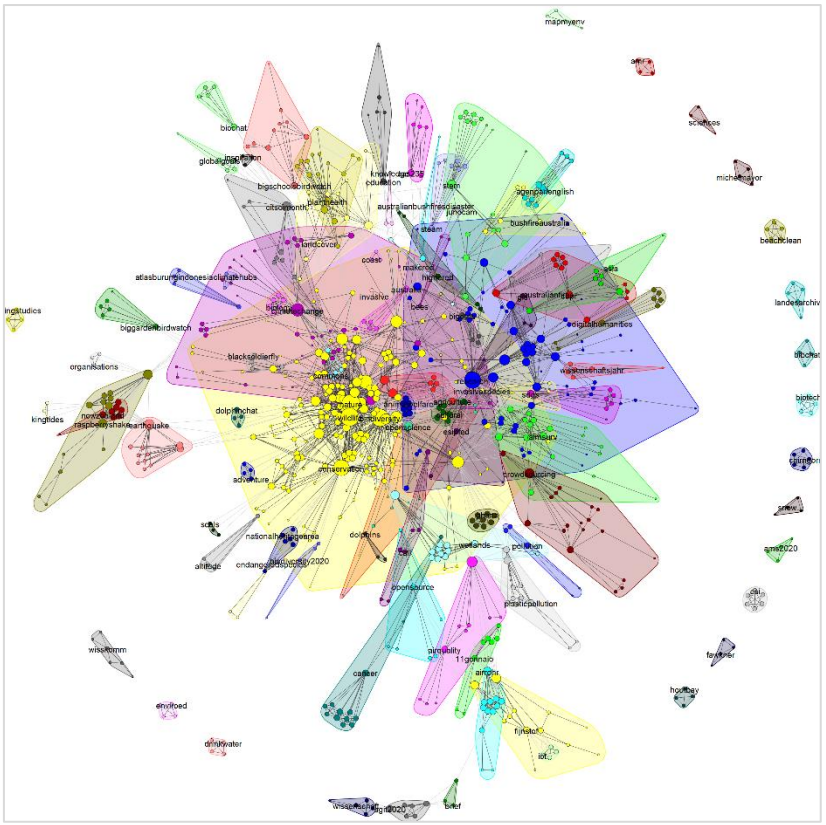

Figura 3-15 Comunidades detectadas mediante Label Propagation con pesos

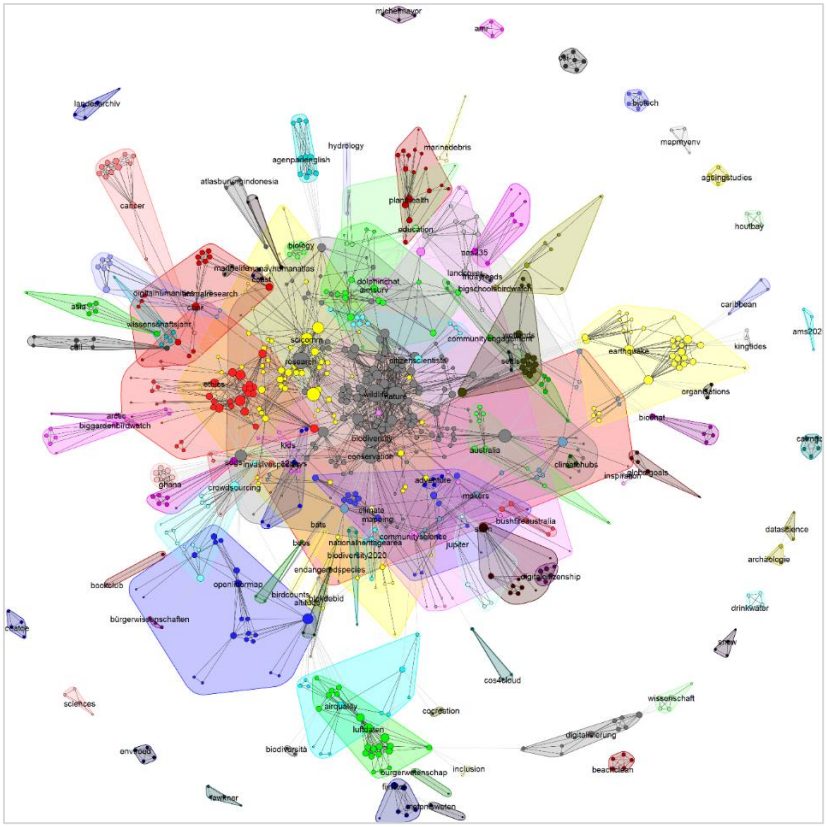

Figura 3-16 Comunidades detectadas mediante Walktrap sin pesos

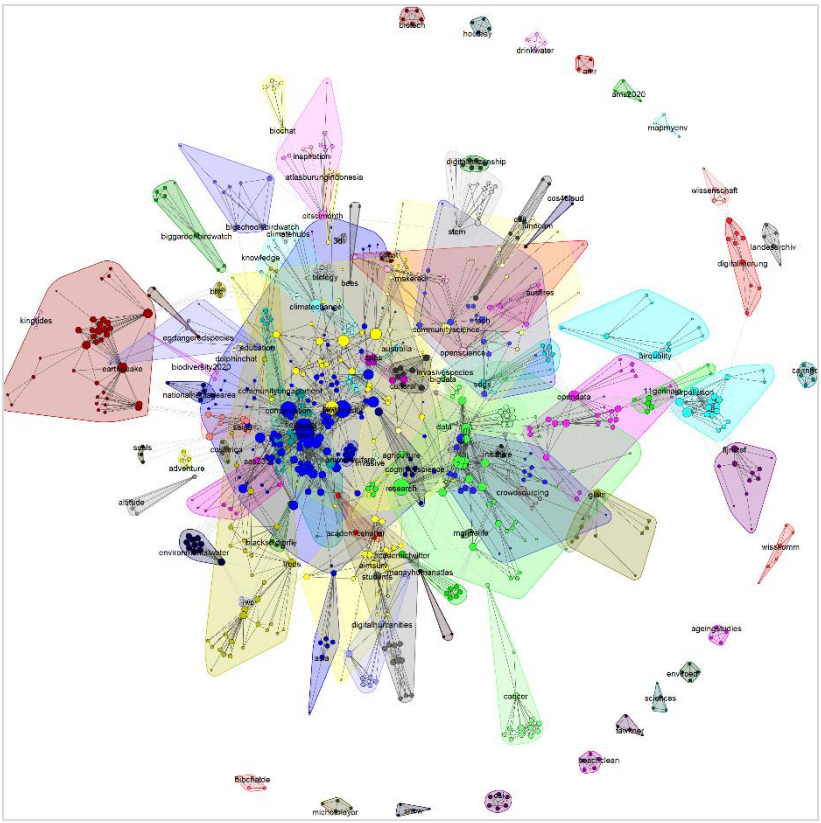

Figura 3-17 Comunidades detectadas mediante Label Propagation sin pesos

Método	*Nº de comunidades*	*Media de hashtags por comunidad*
Edge betweenness	237	20
Walktrap con pesos	336	10
Walktrap sin pesos	196	27
Label propagation con pesos	158	37
Label propagation sin pesos	74	80

Tabla 3-3 Comunidades en función del algoritmo

3.5. Información sobre las comunidades

Toda la información generada con los algoritmos de detección de comunidades se encuentra almacenada en sus correspondientes ficheros en el repositorio de GitHub, con la nomenclatura "InfoComunidadesAlgoritmoGeneración.txt". Se muestra a continuación un caso sencillo sobre la extracción realizada para una de las comunidades:

Comunidad 92:

***Organizaciones**: ['Archaeology' 'FlowerCounts' 'Freo / Walyalup'*

'HeritageFundUK Community Science Project' 'HeritageTreasures' 'JNCC'

'MammalMapperApp' 'Science' 'Sunset/Sunrise' 'TCV Scotland' 'WSTC6'

'the JNCC Bulletin']

***Personas**: ['Chris Stacey' 'Fox' 'Freo' 'Jens Hörig' 'Könnt' 'Stuart' 'Vivian']*

***Proyectos**: ['HeritageFundUK Community Science Project']*

Tras realizar este proceso, se ha detectado, en el caso de Walktrap el siguiente número de entidades y proyectos en total:

Organizaciones	**Personas**	**Proyectos**
4231	2358	143

Tabla 3-4 Número total de entidades y proyectos en Walktrap

Se complementa la información con un cálculo de las diez entidades y proyectos más mencionados:

Organizaciones	Nº	Personas	Nº	Proyectos	Nº
NASA	53	Day	25	EUCitSciProject	12
EU	36	Prof	12	ProjectGreco	6
SciStarter	30	Mark	11	GLOBEProgram	6
UN	24	Hunt	10	Clump Scout Project	5
Twitter	22	Mary Ellen Hannibal	9	PericlesProject	4
Facebook	21	James	8	Anniversary of The GLOBE Program	4
NOAA	18	BirdsCanada	8	the SF City Nature Challenge Project	3
OpenUniversity	16	Vogel	7	the Lost Ladybug Project	3
CitSciAssoc	15	RensvdSchoot	7	the Earth Project	3
BBC	14	Maike Weißpflug	7	the AIMsurv COSTprogramme	3

Tabla 3-5 Organizaciones, personas y proyectos con más ocurrencias

3.6. Presentación de los datos

Para presentar los resultados obtenidos, se ha diseñado un cuadro de mandos en un sitio web, que se encuentra disponible de forma online para su visualización en el enlace:

http://tfmdiegoribas.epizy.com/

Durante el desarrollo se han empleado las siguientes tecnologías:

- **HTML5**: se trata de la quinta revisión de un lenguaje de etiquetas, estandarizado para diseño web, mediante el que estructurar y presentar el contenido de la página. Al interpretar el código, el navegador es capaz de determinar dónde y cómo se deben representar los diferentes elementos; texto, imágenes, etc. También se le otorga un significado a cada contenido, pudiendo dividirse como encabezado, párrafos, secciones, pie de página…
- **CSS3**: es la tercera versión de un lenguaje de diseño gráfico, mediante el cual se establecerá el diseño visual de un documento escrito en un lenguaje de etiquetas, como puede ser HTML. Ambas tecnologías han estado siempre muy ligadas, con el objetivo de diferenciar entre el contenido de la página y su estilo.

Permite establecer los colores, estilo de la letra, tamaño y disposición de cada elemento, etc.

- **JavaScript**: lenguaje de programación orientado a objetos, que permite modificar el contenido de la web de manera dinámica, mejorando la interfaz de usuario. Además, su uso se ha ido extendido, utilizándose a menudo entre otras cosas para el envío y recepción de datos desde el cliente a un servidor. Junto a HTML y CSS, forman las tres capas principales de una página web.
- **Bootstrap** (https://getbootstrap.com/): es un framework empleado en diseño web, que simplifica el desarrollo aportando plantillas para numerosos elementos, menús, navegación, galerías de imágenes, etc. Destaca también por facilitar la relación de páginas web responsive, es decir, que su visualización se adapta y adecúa a todo tipo de dispositivos y tamaños, ya sean ordenadores, tablets, o móviles. Curiosamente, esta biblioteca fue desarrollada por Twitter, inicialmente como medida para estandarizar sus herramientas internas, y poco después pasó a ser liberada como código abierto.
- **Infinity Free** (https://infinityfree.net/)**:** es un servicio en línea que permite la publicación totalmente gratuita de un sitio web. Ofrece un servidor en el que poder subir todos los archivos necesarios para el correcto funcionamiento de la página. En este caso no existe límite de almacenamiento, y presenta una disponibilidad del 99,9% del tiempo.

Se exponen a continuación los componentes principales de la web, mostrando en las figuras bocetos de estos con datos falsos para su simulación.

En la página inicial, se ha incluido una cabecera que contiene el logo de la Universidad Autónoma de Madrid y de la Escuela Politécnica Superior. Ambas imágenes son un enlace a su correspondiente sitio web. Entre ambas, y centrado en el medio, se ha incluido un título y subtítulo.

Figura 3-18 Cabecera del sitio web

Inmediatamente después se ha realizado un contador mediante Bootstrap y JavaScript, que muestra la información sobre el número de tuits, hashtags, usuarios y comunidades tratadas. Este contador se anima de forma dinámica nada más acceder a la web, haciéndose una transición que va sumando valores desde uno hasta su valor final.

Figura 3-19 Boceto de la sección de contadores

Tanto la cabecera como el contador son elementos comunes que se encontrarán fijados, con independencia de que se realice la navegación a otras páginas.

El elemento principal consiste en un menú formado por cuatro botones, cada uno con un icono representativo, que al pulsarlos navegan hasta la sección indicada en su nombre.

Figura 3-20 Menú del panel de control

En las tres primeras secciones, Grafos, Gráficas y Series, se muestran las imágenes de los resultados obtenidos es formato thumbnail o miniatura, que consiste en una imagen de vista previa reducida del original, con su correspondiente título, y que al pinchar sobre ella se ampliará para poder visualizarla detalladamente.

En cada sección se muestra un enlace arriba a la izquierda para regresar a la página principal.

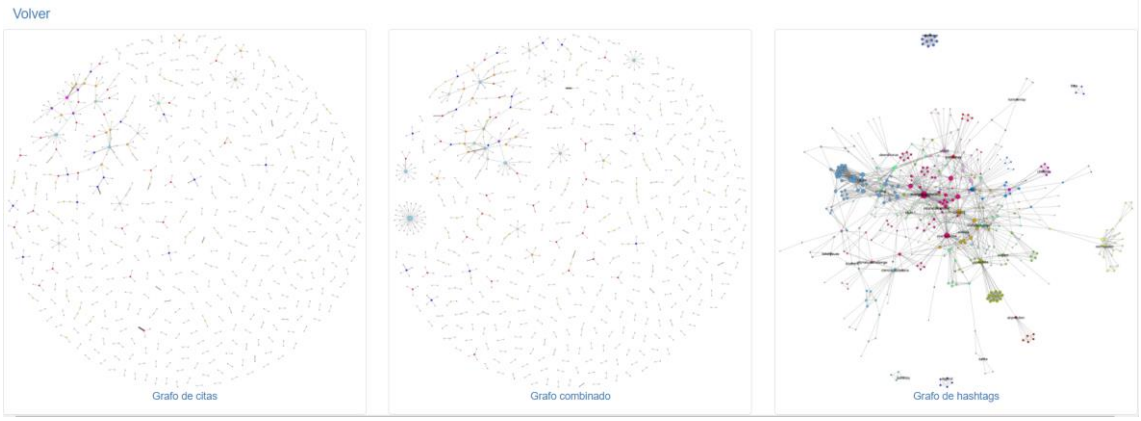

Figura 3-21 Ejemplo de sección con imágenes

3.6 Presentación de los datos

En la última sección, Comunidades, se ha utilizado el patrón de acordeón para mostrar cada una de ellas. Al pulsar sobre el nombre de la comunidad, se despliega su información, y se contrae en el caso de que se vuelva a clicar o se seleccione otra comunidad.

Figura 3-22 Boceto de la sección de comunidades

4. Discusión

El resultado esperado de este trabajo es la detección y categorización de los proyectos de ciencia ciudadana que se generan en Twitter.

Apreciando las componentes conexas en los grafos de la Figura 3-1, Figura 3-2 y y Figura 3-3 se observa que existe una parte importante de la conversación que está conectada, formada generalmente por estructuras con forma de estrella, que a su vez, de forma directa o indirecta, terminan estando conectadas también entre sí. Además de poder observarse a simple vista sobre las gráficas, atendiendo a la distribución de grado de estos grafos expuestas en el Anexo III, es posible concluir que, con independencia de ciertas comunidades o interacciones aisladas, sí existe una comunidad más o menos consolidada, liderada por ciertos usuarios que cuentan con una capacidad de dinamización mayor, cohesionando así la comunidad.

Por otro lado, al utilizar técnicas de NTA y relacionar los hashtags que aparecen en el mismo tuit, en la Figura 3-4, también se muestra claramente que hay intereses comunes (pues existe un componente gigante) a pesar de que no hay una conversación cohesionada. El análisis de las comunidades se realiza sobre esta red de hashtags, donde se identificarán diferentes subgrupos de interés, reforzándose así la idea de que existe una serie de intereses comunes, aunque no interactúen o se conozcan entre sí.

Realizando un análisis de los hashtags más retuiteados que se muestran en la Figura 3-5, como es de esperar por haber sido el acontecimiento más destacado del año, destacan varios hashtags referentes al coronavirus (“covid19”, “coronavirus” y “covid2019”). Predominan también los hashtags relacionados con el medio ambiente (“earthquake”), sobre la biodiversidad (“biodiversity”) y la naturaleza (“nature”), ya que suelen ser uno de los temas más importantes en lo que respecta a la ciencia ciudadana. Se puede destacar por último la aparición del hashtag “nasaathome”, ya que, por los datos observados, parece que la NASA se encuentra bastante involucrada en este tipo de comunidades, y la repercusión que presenta es bastante mayor que la que pueden tener otras cuentas. Se puede comprobar teniendo en cuenta que este hashtag ha sido retuiteado 1653 veces, y se ha usado como hashtag principal 103 veces. Esto implica que cada publicación que contenía este hashtag ha sido retuiteada del orden de 16 veces. Además, como puede observarse en la Tabla 3-5, la NASA es la organización más mencionada entre las distintas comunidades.

En cuanto a los hashtags principales de la Figura 3-6, destacan entre los seis primeros etiquetas relacionadas con la medición de la calidad del aire. Esto se debe a la existencia de un bot que va publicando constantemente los resultados obtenidos de las mediciones utilizando dichos hashtags. Es curioso observar en la Figura 3-11 que, a principios de cada mes, durante unos días no se realizan estas publicaciones, mientras que habitualmente el bot publica un tuit cada 10 minutos, lo que genera la diferencia tan grande que se aprecia en la gráfica. Se desconoce el motivo al respecto, aunque es

posible que se deba a algún tipo de mantenimiento durante esos días. Como la existencia de estos tuits contaminan la comprensión del contexto general de la información, se ha generado también la Figura 3-7 y Figura 3-12, mostrando los mismos datos para los siguientes hashtags, descartando los propios del bot. Con este nuevo enfoque aparecen de nuevo temas recurrentes: coronavirus ("covid19", "coronavirus") y medio ambiente ("biodiversity", "earthquake", "airquality", "seismograph", "nature"). En la nueva gráfica sobre la evolución temporal, cabe destacar el pico producido en la utilización del hashtag "biodiversity" a finales de mayo, ya que el día 22 de mayo es el Día Internacional de la Biodiversidad.

La misma situación sucede al extraer los términos con mayor frecuencia en todo el texto en la Figura 3-9, que podemos ver que en su mayoría hacen referencia a los términos publicados sobre las mediciones del aire.

En la Figura 3-10, analizando la evolución de los hashtags más retuiteados, destaca especialmente el aumento que se produce sobre el coronavirus en los inicios de la pandemia y durante sus puntos más álgidos. Es posible ver que, de forma general, los temas relacionados con la biodiversidad se mantienen más o menos constantes. Sobre el hashtag "nasaathome", si se analizan los datos pueden encontrase dos picos principales coincidiendo con la publicación de ciertos eventos: uno en abril al publicarse una aplicación capaz de detectar corales, y otro en mayo cuando se realizó un evento en vivo sobre proyectos de ciencia ciudadana con expertos de la NASA.

Respecto a la detección de comunidades, dado que la ejecución se está realizando sobre un ordenador personal con una capacidad limitada, se ha descartado la utilización del algoritmo Edge betweenness por requerir de un alto coste computacional, al ser necesario calcular todos los caminos cortos que existen entre dos nodos, lo que le conlleva a ser un método ineficiente para volúmenes elevados de datos. Por otra parte, en los casos de Walktrap y Label propagation parece más interesante utilizar en ambos el peso de las aristas, ya que refuerza la relación entre hashtags de una misma comunidad, y evita agruparlos en los casos en los que ambos hashtags aparezcan juntos de forma prácticamente aislada.

El último análisis consiste en la elección entre ambos algoritmos para determinar por cuál de ellos decantarse. Para ello, se deben analizar las comunidades formadas, tanto por los hashtags que las componen como por las entidades que se han obtenido con el fin de buscar cuáles de ellas presentan una relación mayor. Aunque no es una tarea fácil, exacta y que puede resultar algo subjetiva, en esta ocasión parece que Walktrap arroja mejores resultados. Mientras que Label Propagation forma las primeras comunidades abarcando un número bastante grande de nodos y con temáticas demasiado diversas, en el caso de Walktrap se aprecia que parecen estar mejor divididas. Por ejemplo: la comunidad 8 cuya temática se centra en la ornitología, la comunidad 24 que trata sobre anfibios, o la comunidad 33 relacionada con el espacio.

Además, estos datos permiten confirmar los temas principales que tratan de abordar y que despiertan un mayor interés en este tipo de comunidades. Entre otras, destacan por

ejemplo la ecología (contaminación atmosférica, preservación de la biodiversidad, jardinería), climatología (terremotos, incendios, tormentas), biología (flora y fauna, enfermedades), o física y tecnología: (exploración espacial, digitalización, smartcities). En el sitio web pueden consultarse todas las comunidades con sus hashtags, proyectos y entidades.

Durante este ejercicio, se presentan ciertas limitaciones. Hay que destacar, entre otras, que las técnicas de lenguaje de procesamiento natural, y en este caso concreto sobre NER a través de la librería Stanza, a menudo no disponen de la suficiente inteligencia y se producen ciertos errores a la hora de extraer y categorizar las entidades. Estos errores se incrementan aún más al tratarse de texto publicado en una red social, en el que los usuarios utilizan a menudo abreviaturas, se cometen faltas ortográficas, y el uso correcto del lenguaje, como signos de puntuación o el empleo correcto de mayúsculas y minúsculas, es más reducido. También influye notablemente el hecho de que en el conjunto de datos con el que se trabaja es posible encontrar publicaciones en diferentes idiomas, lo que aumenta su complejidad.

5. Conclusiones y trabajo futuro

5.1. Conclusiones

Para lograr el objetivo principal se ha implementado un sistema informático que caracteriza los proyectos de ciencia ciudadana a partir de los datos publicados en Twitter. Para llevarlo a cabo, se destacan la consecución de las siguientes fases:

- Se ha realizado la extracción, comprensión, preprocesamiento e ingesta de los datos que han servido como fuente para la realización del trabajo.
- Se han analizado las diferentes formas en las que relacionar la información de la que se dispone y distintos algoritmos para conseguir la obtención de las comunidades existentes sobre ciencia ciudadana.
- Utilizando minería de textos, se ha contextualizado y caracterizado la información, detectando los temas más abundantes en ella, así como entre las distintas interacciones de los usuarios, mostrando en ambos casos la evolución temporal de cada uno de ellos.
- Se ha enriquecido la información de cada comunidad mediante el uso de técnicas de procesamiento de lenguaje natural, complementando los hashtags que forman dicha comunidad con las organizaciones, personas y proyectos implicados en las misma.
- Los resultados se han expuesto a través de un cuadro de mandos que facilita su visualización, encontrándose disponibles para su consulta online.

Además de cumplirse con los objetivos planteados, en lo personal este trabajo ha supuesto un reto dado el escaso conocimiento inicial tanto en el ámbito del proyecto, como en el lenguaje de programación empleado, pero a su vez conlleva una mayor satisfacción por la superación de estas dificultades, y por los nuevas conocimientos y habilidades adquiridas durante el proceso.

5.2. Trabajo futuro

Debido a la limitación tanto temporal como de recursos para la realización de este proyecto, se plantean a continuación algunos aspectos de mejora y evolución que podrían aplicarse:

- **Rendimiento**: realizar una revisión completa del código y el proceso para detectar posibles optimizaciones de tiempo. También sería interesante plantear

la ejecución de este sobre un clúster de ordenadores, potenciando así la capacidad de procesamiento.

- **Algoritmos de detección de comunidades**: se trata de un mundo complejo y muy extenso. Es posible realizar un estudio más en profundidad, incluyendo nuevos algoritmos diferentes a los empleados, y analizando en detalle cuál sería el más adecuado en este caso.
- **Procesamiento del lenguaje natural**: investigar y probar el uso de diferentes librerías a Stanza para la extracción de entidades.
- **Visualización de los resultados:** mejorar el diseño y usabilidad de la página web para mostrar los resultados de la forma más adecuada para el usuario y mejorar la estética de la página.

Referencias

Agarwal, N., Kumar, S., Gao, H., Zafarani, R., Liu, H. (2012). Analyzing behavior of the influentials across social media. In: Behavior Computing (pp. 3-19). Springer Nature.

Borgatti, S. et al. (2009) Network Analysis in the Social Sciences, Science, 323: 892

Carley, K. M. (1997). Extracting team mental models through textual analysis. Journal of Organizational Behavior, 533-558.

Carley, K., & Palmquist, M. (1992). Extracting, representing, and analyzing mental models. Social Forces, 70(3), 601-636.

Dubitzky W., Wolkenhauer O., Cho KH., Yokota H. (2013) In: Edge Betweenness. Encyclopedia of Systems Biology. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9863-7_101489

Hoppe, H.U. (2017). Computational methods for the analysis of learning and knowledge building communities. In Lang, C., Siemens, G., Wise, A., Gasevic, D. (2017), Handbook of Learning Analytics, Chapter 2 (pp. 21-40). Available under https://solaresearch.org/hla-17/#

Grafo. (s.f.). En Wikipedia. Recuperado el 12 de agosto de 2020 de https://es.wikipedia.org/wiki/Grafo

Leskovec, J., Backstrom, L., Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In: Proc. of the 15th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining (pp. 497-506). ACM.

Pons, P. & Latapy. (2005). M. Computing communities in large networks using random walks. In Computer and Information Sciences-ISCIS 2005, 284–293, Springer

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. and Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations.

Wasserman and Faust. (1994). Social Network Analysis, Cambridge University Press, Cambridge

X. Zhu, Z. Ghahramani. (2002), Learning from Labeled and Unlabeled Data with Label Propagation, Technical Report CMU-CALD-02-107, School of Computer Science, Carnegie Mellon University.

Bibliografía

Bonchi, F., Castillo, C., Gionis, A., and Jaimes, A. Social network analysis and mining for business applications. ACM Transactions on Intelligent Systems and Technology 2, 3 (Apr. 2011), 22.

Cha, M., Haddadi, H., Benevenuto, F. & Gummadi, K.P (2010) Measuring user influence in Twitter: the million follower fallacy. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM 2010) (pp. 10-17)

Eytan Bakshy, Jake M Hofman, Duncan J Watts, Winter A Mason. Identifying Influencers on Twitter ACM International Conference on Web Search and Data Mining (WSDM), 2011

Harry Bitten (2019) Detecting communities in a language co-occurrence network. Towardsdatascience. Recuperado de https://towardsdatascience.com/detecting-communities-in-a-language-co-occurrence-network-f6d9dfc70bab

Fortunato, S. (2010) Community Detection in Graph. Physics Reports, 486, 75-174. http://dx.doi.org/10.1016/j.physrep.2009.11.002

Goel, S., Watts, D. J., & Goldstein, D. G. (2012) The structure of online diffusion networks. In Proceedings of the 13th ACM Conference on Electronic Commerce (EC '12). ACM, New York, NY, USA, 623-638. http://dx.doi.org/10.1145/2229012.222905

Kwak, H, Lee, C, Park, H. & Moon, S (2010) What is Twitter, a social network or a news media?. In Proceedings of the 19th international conference on World wide web (WWW '10). ACM, New York, NY, USA, 591-600. http://dx.doi.org/10.1145/1772690.1772751

Scott, J. Social Network Analysis. SAGE Publications, Thousand Oaks, CA, 2012.

Zeng, D., Chen, H., Lusch, R., and Li, S-H. Social media analytics and intelligence. IEEE Intelligent Systems 25, 6 (Dec. 2010), 13–16.

Anexo I: Código fuente y ficheros

El código fuente utilizado y los ficheros relativos al proyecto pueden encontrase en el siguiente repositorio de GitHub:

https://github.com/diegoribas/TFMUAM

El repositorio se encuentra estructurado en tres partes:

- Grafos.py: es el fichero que contiene todo el código fuente de Python llevado a cabo.
- Archivos: carpeta que contiene los archivos con los datasets y los ficheros generados al exportar la información de las comunidades.
- Web: código fuente de la página realizada para mostrar de forma online los resultados a través de un panel de control.

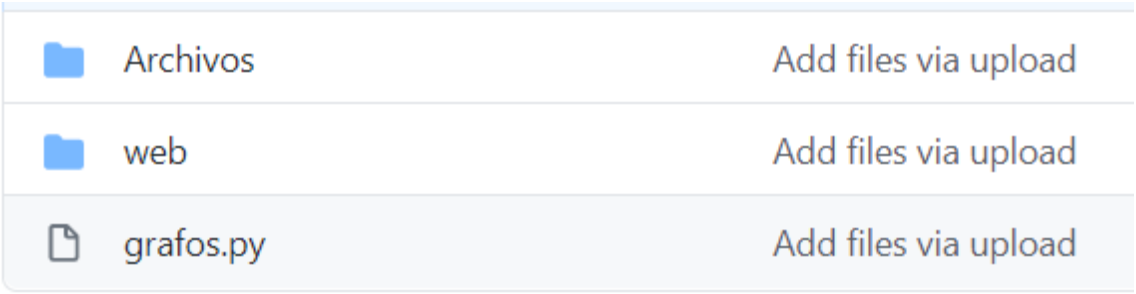

Figura I-1 Estructura del repositorio en GitHub

Anexo II: Grafos completos

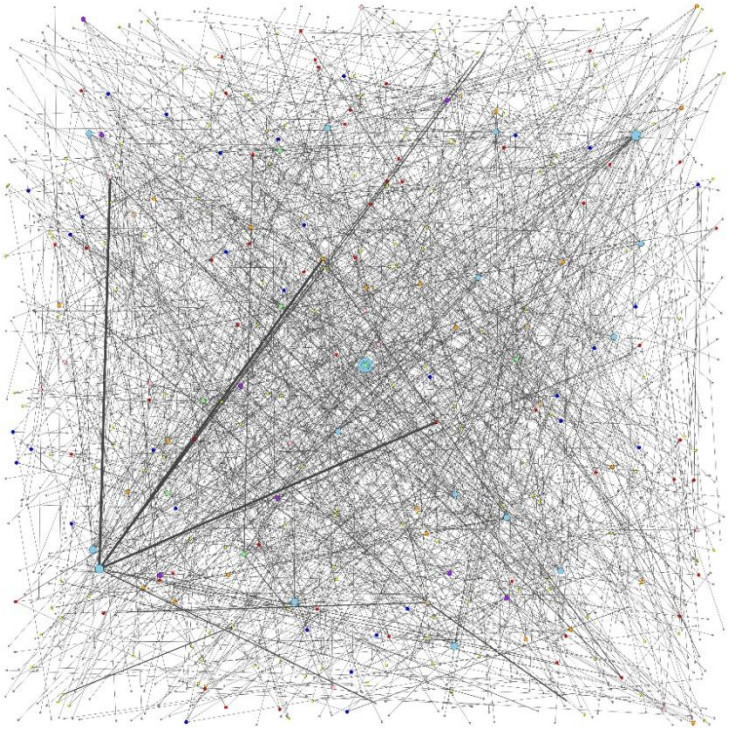

Figura II-1 Grafo completo de tuits

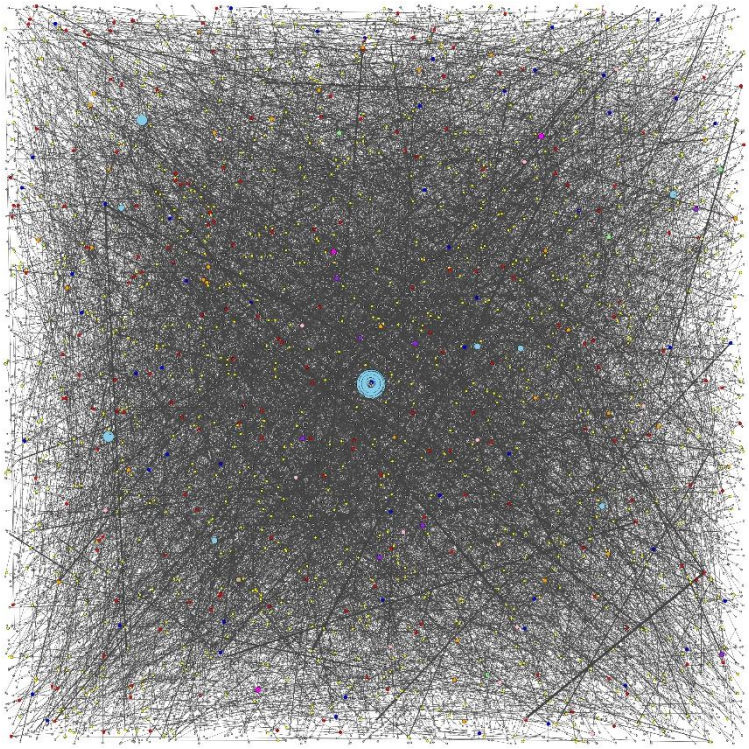

Figura II-2 Gráfica completo de citas

Anexo III: Distribuciones de grado

Grafo de retuits:

- Número de elementos: 2.686
- Media: 1,5972
- Desviación estándar: 2,7269

Grado	Número de elementos
0	119
1	2040
2	233
3	111
4	59
5	34
6	25
7	11
8	12
9	9
10	1
11	3
12	5

Grado	Número de elementos
13	4
14	3
16	3
17	1
18	3
19	3
22	1
26	2
29	1
34	1
57	1
74	1

Tabla III-1 Distribución de grado del grafo de retuits

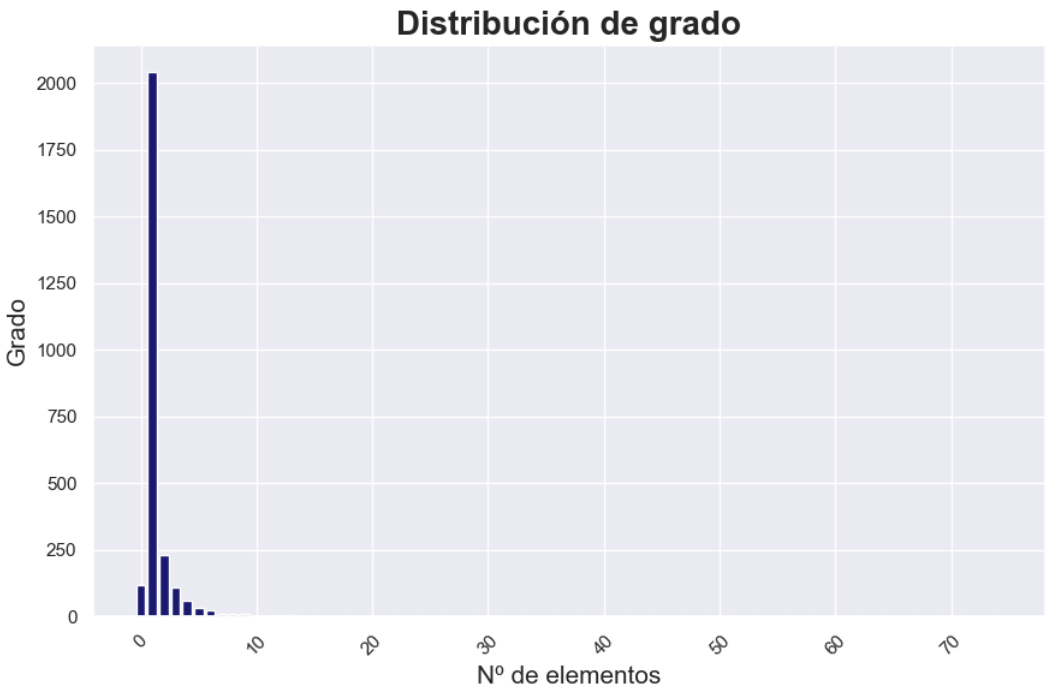

Figura II-3 Gráfica distribución de grado en retuits

Grafo de citas:

- Número de elementos: 21.524
- Media: 1,795
- Desviación estándar: 4,8857

Grado	Número de elementos
0	84
1	16433
2	2666
3	888
4	417
5	244
6	157
7	137
8	101
9	64
10	52
11	40
12	36
13	25
14	18
15	8
16	16
17	8
18	11
19	8
20	8
21	10
22	3

Grado	Número de elementos
23	7
24	4
25	5
26	1
27	5
28	2
29	4
30	3
31	2
32	2
33	1
34	3
35	2
37	3
38	1
39	7
40	4
41	1
45	1
46	2
47	2
48	1
49	2

Grado	Número de elementos
50	1
54	1
56	1
58	1
59	2
64	1
65	1
66	2
68	1
77	2
78	2
84	1
95	1
109	1
116	1
123	1
196	1
200	1
206	1
270	1
275	1

Tabla III-2 Distribución de grado del grafo de citas

Figura II-4 Gráfica distribución de grado en citas

Grafo combinado:

- Número de elementos: 22.799
- Media: 1.8684
- Desviación estándar: 5.1308

Grado	Número
1	17103
2	2898
3	1006
4	499
5	295
6	179
7	159
8	111
9	61
10	62
11	45
12	48
13	35
14	23
15	16
16	23
17	9
18	15
19	7
20	7
21	8
22	6
23	6
24	5
25	7
26	1
27	4
28	2
29	1
30	3
31	4
32	5
33	2
34	3

Grado	Número
35	3
36	3
38	1
39	6
40	3
41	3
42	0
46	1
47	1
48	1
49	0
50	0
51	2
55	1
57	1
59	2
60	0
65	0
66	2
67	0
69	1
78	0
79	1
85	1
96	0
110	0
117	0
124	0
197	0
201	0
207	0
271	0
276	0
277	0

Grado	Número
278	0
279	0
280	0
281	0
282	0
283	0
284	0
285	0
286	0
287	0
288	0
289	1
290	0
291	0
292	0
293	0
294	0
295	0
296	0
297	0
298	0
299	0
300	0
301	0
302	0
303	0
304	0
305	0
306	0
307	0
308	0
309	1

Tabla III-3 Distribución de grado en grafo combinado

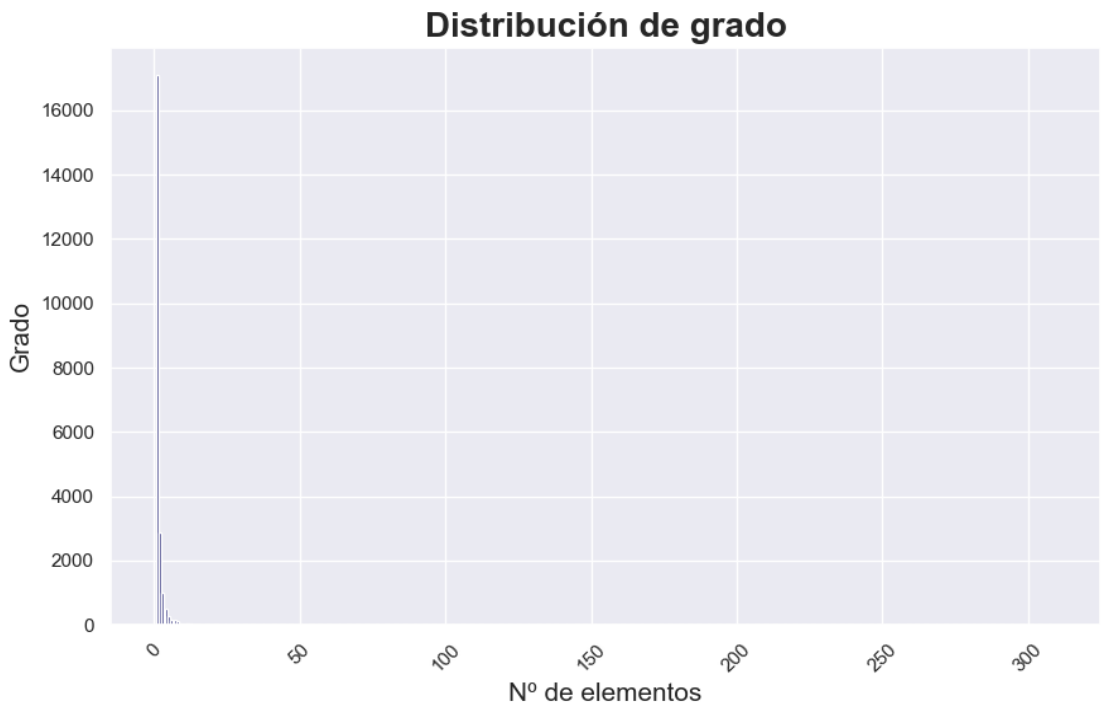

Figura II-5 Distribución de grado del grafo combinado